

רגרסיה ומודלים לינאריים

מבוסס על הרצאות פרופ' אור צוק
בקורס "רגרסיה ומודלים לינאריים" (52320)
האוניברסיטה העברית, ספטמבר ב' 2015
להערות: *nachi.avraham@gmail.com*

נחי

תודה למי ששלח הערות ותיקונים: מתן כהן, דן נבון, רבקה סואר

תוכן עניינים

4	I רגרסיה לינארית פשוטה	
4	1 מבוא	
5	2 רגרסיה לינארית פשוטה - ישר הריבועים הפחותים	
8	2.1 סטטיסטיים נפוצים	
9	2.2 תכונות ישר הריבועים הפחותים	
11	3 רגרסיה לינארית פשוטה - מודל סטטיסטי כללי לשגיאות	
11	3.1 תוחלות של האומדים	
13	3.2 שונויות של האומדים	
15	3.2.1 פירוק השונות לרכיבים	
18	4 רגרסיה לינארית פשוטה - מודל סטטיסטי נורמלי לשגיאות	
18	4.1 אומדי נראות מקסימלית	
19	4.2 הסקה סטטיסטית	
19	4.2.1 רווח סמך של β_1	
21	4.2.2 בדיקת השערות על β_1	
22	II רגרסיה לינארית מרובת משתנים	
22	5 מבוא	
	5.1 רגרסיה לינארית מרובת משתנים - הצגה מטריציונית ומודל סטטיסטי נורמלי	
23	6 שיטת הריבועים הפחותים	
24	6.1 הוכחה ראשונה: באמצעות גזירה	
24	6.1.1 מבוא: מושגים כלליים באינפי של כמה משתנים	
24	6.1.2 הוכחת משפט הריבועים הפחותים	
26	6.2 הוכחה שנייה: באמצעות הטלות	
26	6.2.1 מבוא: הטלות	
26	6.2.2 הוכחת משפט הריבועים הפחותים	
27	6.3 הסקה סטטיסטית	
27	6.3.1 התפלגות $\hat{\beta}$	
28	6.3.2 אמידת σ^2	
30	6.3.3 רווח סמך של רכיבי הווקטור β	
30	6.3.4 בדיקת השערות על רכיבי הווקטור β	
30	7 טרנספורמציות	
32	8 משתנים קטגוריים	
32	9 פירוק לסכום ריבועים	
34	10 הסקה סטטיסטית - השוואת מודלים מקוננים	
39	10.1 מבחן F מול מבחן t	
40	11 Adjust R^2	
40	12 בדיקת הנחות המודל	
41	13 ייצוב השונות	
44	14 מדדים לחריגות של תצפית	
44	14.1 הנפה	
45	14.2 מרחק קוק	

46	14.2.1	התפלגות מרחק קוק	
47	15	מדדים למולטי-קוליניאריות של משתנה	
48	16	אינטראקציות	
49	17	בחירת מודל	
50	17.1	מדידת טיב של מודל	
50	17.1.1	שגיאת התחזיות הצפויה (Expected Prediction Error)	

חלק I

רגרסיה לינארית פשוטה

1 מבוא

נתונים מספרים ממשיים $X_1, \dots, X_n, Y_1, \dots, Y_n$. נתייחס אליהם כזוגות $(X_1, Y_1), \dots, (X_n, Y_n)$. כל זוג כזה נכנה **תצפית**. המטרה המרכזית היא "להסביר" את הנתונים $Y_1, \dots, Y_n$ באמצעות הנתונים $X_1, \dots, X_n$. בהתאם לכך, נקרא לנתונים $X_1, \dots, X_n$ "משתנים מסבירים" ולנתונים $Y_1, \dots, Y_n$ "משתנים מוסברים".

לצורך כך נחפש קשר בין המשתנים X_i, Y_i עבור $i = 1, \dots, n$. "קשר" הוא מודל בדמות פונקציה f המקיימת $f(X_i) \approx Y_i$. תוך שימוש במודל f , נוכל עבור X חדש לנבא את ערך ה- Y המתאים לו, על ידי $Y = f(X)$.

דוגמה: נניח כי $X_1, \dots, X_{365}$ הם ימי השנה. נניח כי $Y_1, \dots, Y_{365}$ הן הטמפרטורות שנמדדו במהלך ימי השנה. כלומר התצפית (X_i, Y_i) היא זוג של היום i -ה' בשנה והטמפרטורה שנמדדה בו.

מודל שקושר בין המשתנה המסביר של היום והמשתנה המוסבר של הטמפרטורה, הוא פונקציה f שבאמצעותה נוכל לנבא בשנה הבאה מה תהיה הטמפרטורה ביום כלשהו.

נרצה כמובן מודל "טוב", כלומר מודל כזה שהסטיות מהערכים האמיתיים יהיו מינימליות. נזכור כי המודל מהווה רק קירוב $f(X_i) \approx Y_i$, ולכן נרצה שבמובן מסוים שנסביר במדויק בהמשך, "המרחק" בין $f(X_i)$ לבין Y_i יהיה מינימלי.

נשים לב שעקרונית ניתן היה לבנות מודל לכאורה מושלם, על ידי הגדרת $f(X_i) = Y_i$ עבור כל $i = 1, \dots, n$, ולהשלים את הגרף של הפונקציה בין הנקודות הללו באופן מלאכותי, למשל על ידי קווים לינאריים. כלומר, במקום לדרוש רק קירוב $f(X_i) \approx Y_i$ נוכל לקבל שוויון ממש, ובכך לייצר את המודל המושלם לכאורה שהסטייה שלו הן פשוט 0.

הבעיה היא שבחירת מודל שכזה היא מוטת לטובת המדגם המסוים שיש לנו. לא נראה זאת באופן פורמלי בקורס, אולם קל לצייר את הגרף ולהבין שפונקציה שנראית באופן כזה היא ניבוי לא טוב של Y על ידי X . תופעה זו מכונה **Overfitting**, וכדי לבנות מודל טוב באמת קיימות שתי גישות כלליות:

הגישה הפרמטרית:¹ נבחר משפחה מסוימת של מודלים שרק מתוכם נבחר את המודל האידיאלי. כלומר, נדרוש שהמודל הקושר בין X לבין Y יהיה דווקא מצורה מסוימת (למשל פולינום ממעלה מסוימת כלשהי), ורק תחת המגבלה הזו נחפש את המודל הטוב ביותר.

הגישה האי-פרמטרית: נבצע "החלקה" של המודל. כלומר נקבע מספר k כלשהו. כעת נבצע את התהליך הבא, עבור כל $i = 1, \dots, n$:

עבור תצפית (X_i, Y_i) נבחר k תצפיות $(X_{i_1}, Y_{i_1}), \dots, (X_{i_k}, Y_{i_k})$ מתוך המדגם שלנו, כך שאוסף k המרחקים בין כל אחד מ- $X_{i_1}, \dots, X_{i_k}$ לבין X_i הם הקטנים ביותר משאר n המרחקים, ואז נגדיר את המודל $f(X_i) = \frac{1}{k} \sum_{j=1}^k Y_{i_j}$.

בקורס הנוכחי נדון רק בגישה הראשונה.

¹בפרק הבא נבין מדוע גישה זו מכונה "פרמטרית".

2 רגרסיה לינארית פשוטה - ישר הריבועים הפחותים

רקע: נתון לנו מדגם של תצפיות $(X_1, Y_1), \dots, (X_n, Y_n)$, כולן זוגות של מספרים ממשיים, ומעוניינים במודל f שקושר בין המשתנה המסביר X_i לבין המשתנה המוסבר Y_i .

נצטמצם למשפחת המודלים הלינאריים. כלומר נחפש מודל f שהוא מהצורה $f(x) = \beta_0 + \beta_1 x$, כאשר β_0, β_1 הם מספרים ממשיים.

לכל מודל לינארי אפשרי f , נגדיר את **השגיאה הריבועית** שלו להיות $(f(X_i) - Y_i)^2$. לפיכך המודל הלינארי שנחפש הוא מודל לינארי שעבורו סכום כל השגיאות הריבועיות $\frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2$ יהיה מינימלי מבין כל שאר המודלים הלינאריים.

סימון: באופן כללי, נסמן ממוצע של מספרים $a_1, \dots, a_m$ על ידי $\bar{a} = \frac{1}{m} \sum_{j=1}^m a_j$.

משפט: המודל הלינארי שעבורו סכום השגיאות הריבועיות הוא מינימלי, הוא המודל הלינארי מהצורה של $f(x) = \beta_0^* + \beta_1^* x$, עבור הערכים:

$$\beta_1^* = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\beta_0^* = \bar{Y} - \beta_1^* \bar{X}$$

הוכחה: אנו מחפשים את המודל שעבורו סכום השגיאות הריבועיות הוא מינימלי, כלומר אנחנו מחפשים את זוג המספרים המקיימים:

$$(\beta_0^*, \beta_1^*) = \arg \min_{\beta_0, \beta_1} \left\{ \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 \right\}$$

נשים לב כי $f(X_i) = \beta_0 + \beta_1 X_i$, ולכן נוכל לכתוב:

$$(\beta_0^*, \beta_1^*) = \arg \min_{\beta_0, \beta_1} \left\{ \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i)^2 \right\}$$

נגדיר פונקציה של שני משתנים, $G: \mathbb{R}^2 \rightarrow \mathbb{R}$ על ידי:

$$G(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i)^2$$

נשים לב שמה שאנחנו מחפשים הוא זוג ערכים (β_0^*, β_1^*) שמביא את G למינימום.

הטכניקה שבה נמצא את זוג הערכים הללו תכלול שני שלבים: (1) תחילה נראה שאם יודעים מהו β_1^* אז ניתן להסיק מהו β_0^* , (2) נמצא את β_1^* .

²ההצדקה לשימוש בטכניקה זו היא שבאופן כללי, כדי למצוא פונקציה של שני משתנים ניתן למצוא כל משתנה בנפרד. כלומר ניתן למצוא קודם לפי משתנה אחד ואחר כך לפי המשתנה השני, לא משנה באיזה סדר.

1. נקבע β_1 כלשהו, ונגדיר פונקציה חדשה של משתנה אחד, $G_1 : \mathbb{R} \rightarrow \mathbb{R}$ על ידי $G_1(\beta_0) = G(\beta_0, \beta_1)$. נשים לב שזו פונקציה של משתנה אחד בלבד, שכן את β_1 קבענו.

כדי למזער את G_1 יש לגזור ולהשוות לאפס. נסמן $\tilde{\beta}_0(\beta_1) = \arg \min_{\beta_0} \{G_1(\beta_0)\}$ ונקבל שמתקיים:

$$\frac{\partial G_1(\beta_0)}{\partial \beta_0} = 2 \cdot \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i) \doteq 0$$

$\Downarrow$

$$\frac{1}{n} \sum_{i=1}^n \beta_0 + \frac{1}{n} \sum_{i=1}^n \beta_1 X_i - \frac{1}{n} \sum_{i=1}^n Y_i \doteq 0$$

$\Downarrow$

$$\beta_0 + \beta_1 \bar{X} - \bar{Y} \doteq 0$$

$\Downarrow$

$$\tilde{\beta}_0(\beta_1) = \bar{Y} - \beta_1 \bar{X}$$

אם כך מצאנו שאם נדע מהו β_1^* , ניתן למצוא את $\beta_0^* = \tilde{\beta}_0(\beta_1^*) = \bar{Y} - \beta_1^* \bar{X}$ ונקבל זוג (β_0^*, β_1^*) שמהווה נקודת קיצון של G .

2. כעת נמצא את β_1^* . תחילה נציב את הערך β_0^* שמצאנו, ונשים לב שמתקיים:

$$G(\beta_0^*, \beta_1) = \frac{1}{n} \sum_{i=1}^n (\beta_0^* + \beta_1 X_i - Y_i)^2 = \frac{1}{n} \sum_{i=1}^n (\bar{Y} - \beta_1 \bar{X} + \beta_1 X_i - Y_i)^2$$

ולכן נפטרנו מהמשתנה β_0 . נגזור לפי β_1 ונשווה לאפס:

$$\frac{\partial G(\beta_0^*, \beta_1)}{\partial \beta_1} = 2 \cdot \frac{1}{n} \sum_{i=1}^n (\bar{Y} - \beta_1 \bar{X} + \beta_1 X_i - Y_i) (X_i - \bar{X}) =$$

$$= 2 \cdot \frac{1}{n} \sum_{i=1}^n (\bar{Y} + \beta_1 (X_i - \bar{X}) - Y_i) (X_i - \bar{X}) \doteq 0$$

$\Downarrow$

$$\frac{1}{n} \sum_{i=1}^n (\bar{Y} - Y_i) (X_i - \bar{X}) + \beta_1 \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \doteq 0$$

$\Downarrow$

$$\beta_1^* = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

כעת עלינו לוודא שנקודת הקיצון שמצאנו (β_0^*, β_1^*) היא נקודת מינימום. לצורך כך נזכר כיצד מאפיינים נקודות קיצון של פונקציה בשני משתנים.

איפיון נקודות קיצון של פונקציה בשני משתנים

סימון: בהינתן פונקציה של שני משתנים $f(x, y)$, היעקוביאן שלה הוא וקטור הנגזרות הראשונות, וההסיאן שלה הוא מטריצת הנגזרות השניות. כלומר, אם נסמן את היעקוביאן J_f ואת ההסיאן H_f , מתקיים:

$$J_f(x, y) = (f_x(x, y), f_y(x, y)) \quad H_f(x, y) = \begin{pmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{pmatrix}$$

הערה: כאשר הנגזרות החלקיות רציפות, מתקיים $f_{xy} = f_{yx}$, ולכן במקרה זה ההסיאן היא מטריצה סימטרית.

תזכורת: בהינתן (x^*, y^*) נקודת קיצון של $f(x, y)$, כלומר $J_f(x^*, y^*) = 0$, אז ניתן לסווג אותה כנקודת מינימום או מקסימום באופן הבא:

אם $\det(H_f(x^*, y^*)) > 0$ וגם $f_{xx}(x^*, y^*) > 0$ אז (x^*, y^*) נקודת מינימום מקומי.
 אם $\det(H_f(x^*, y^*)) > 0$ וגם $f_{xx}(x^*, y^*) < 0$ אז (x^*, y^*) נקודת מקסימום מקומי.
 אם $\det(H_f(x^*, y^*)) = 0$ אז לא ניתן לדעת בהסתמך רק על נתונים אלה האם (x^*, y^*) היא מינימום מקומי, מקסימום מקומי או נקודת אוכף.

$$G(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i)^2$$

נחשב תחילה את היעקוביאן:

$$J_G(\beta_0, \beta_1) = (G_{\beta_0}(\beta_0, \beta_1), G_{\beta_1}(\beta_0, \beta_1)) =$$

$$= \left(2 \cdot \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i), 2 \cdot \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i) \cdot X_i \right)$$

וכעת נחשב את ההסיאן:

$$H_G(\beta_0, \beta_1) = \begin{pmatrix} G_{\beta_0\beta_0}(\beta_0, \beta_1) & G_{\beta_0\beta_1}(\beta_0, \beta_1) \\ G_{\beta_1\beta_0}(\beta_0, \beta_1) & G_{\beta_1\beta_1}(\beta_0, \beta_1) \end{pmatrix} = \begin{pmatrix} 2n & 2 \cdot \sum_{i=1}^n X_i \\ 2 \cdot \sum_{i=1}^n X_i & 2 \cdot \sum_{i=1}^n X_i^2 \end{pmatrix}$$

מצאנו כי (β_0^*, β_1^*) נקודת קיצון. נחשב את הדטרמיננטה של ההסיאן בנקודה זו:

$$\det(H_G(\beta_0^*, \beta_1^*)) = 4n \cdot \sum_{i=1}^n X_i^2 - 4 \left(\sum_{i=1}^n X_i \right)^2 =$$

$$4n \left[\sum_{i=1}^n X_i^2 - n \cdot \bar{X}^2 \right] = 4n \sum_{i=1}^n (X_i - \bar{X})^2 > 0$$

כאשר השוויון השני נובע מכך שמתקיים $(\sum_{i=1}^n X_i)^2 = n^2 \cdot \bar{X}^2$.

אם כך קיבלנו כי ההסיאן חיובי (אפילו ללא תלות ב- (β_0^*, β_1^*) עצמה), ולכן בהתאם לתזכורת שהבאנו נובע כי (β_0^*, β_1^*) נקודת מינימום. ■

2.1 סטטיסטיים נפוצים

תזכורת: בהינתן מדגם כלשהו $Z_1, \dots, Z_n$, סטטיסטי של המדגם הוא פונקציה כלשהי מהצורה $g(Z_1, \dots, Z_n)$. כלומר, זו פונקציה התלויה אך ורק במדגם ולא באף פרמטר אחר (למשל לא בפרמטר של ההתפלגות ממנה מגיע המדגם).

להלן נגדיר סטטיסטיים נפוצים המשמשים ברגרסיה לינארית. נתייחס למדגם נתון מהצורה $(X_1, Y_1), \dots, (X_n, Y_n)$, ונגדיר את הסטטיסטיים הבאים:

• **סטיית תקן אמפירית:**

$$s_X := \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$s_Y := \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

• **מקדם המתאם (של פירסון):**

$$r_{X,Y} := \frac{\overline{X \cdot Y} - \bar{X} \cdot \bar{Y}}{\sqrt{\overline{X^2} - \bar{X}^2} \sqrt{\overline{Y^2} - \bar{Y}^2}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

כאשר השוויון השני הוא זהותי.

תזכורת: מקדם המתאם אינו משתנה בעקבות טרנספורמציה לינארית ששומרת סימן.

כלומר, לכל $a \geq 0$ ולכל b , מתקיים כי $r_{aX+b,Y} = r_{X,Y}$.

נשים לב כי $r_{X,Y}$ סימטרי ב- X, Y , ולכן מידית גם $r_{X,aY+b} = r_{X,Y}$.

• **סכומים שונים:**

$$S_X := \sum_{i=1}^n X_i$$

$$S_Y := \sum_{i=1}^n Y_i$$

$$S_{XX} := \sum_{i=1}^n X_i^2$$

$$S_{YY} := \sum_{i=1}^n Y_i^2$$

$$S_{XY} = S_{YX} := \sum_{i=1}^n X_i Y_i$$

• **שגיאה (residual):** תחילה לכל $i = 1, \dots, n$ נסמן $\hat{Y}_i = \beta_0^* + \beta_1^* X_i$. כעת לכל $i = 1, \dots, n$ נגדיר:

$$e_i := Y_i - \hat{Y}_i$$

כלומר, השגיאה מוגדרת להיות ההפרש בין הערך של Y_i שנצפה במדגם, לבין הערך של $\beta_0^* + \beta_1^* X_i$ שהוא הערך החזוי.

• סכומי ריבועים:

$$SST := \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SSE := \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

הסטטיסטי SST מודד את המרחקים שבין הנתונים לבין הממוצע שלהם, והסטטיסטי SSE מודד את המרחקים שבין הנתונים לבין הערך החזוי שלהם.

טענה: תמיד $SSE \leq SST$.

הוכחה: הגדרנו לעיל את הפונקציה $G(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i)^2$ לא קשה לראות כי $G(0, \bar{Y}) = SST$ וכי $G(\beta_0^*, \beta_1^*) = SSE$ (סימנו $\hat{Y}_i = \beta_0^* + \beta_1^* X_i$). נזכור כי (β_0^*, β_1^*) הוא זוג הערכים שממזער את G , ולכן בהכרח $SSE = G(\beta_0^*, \beta_1^*) \leq G(0, \bar{Y}) = SST$ ■

• פרופורציית השונות המוסברת:

$$R^2 := 1 - \frac{SSE}{SST}$$

נשים לב שמהטענה הקודמת נובע כי $R^2 \leq 1$.

סטטיסטי זה של המדגם מודד את החוזק של הקשר שבין X_i לבין Y_i .

נשים לב כי אם $SSE = 0$, כלומר כל ה- Y_i שווים, אזי ישר הריבועים הפחותים הוא הישר $y = \bar{Y}$ (כי הערכים $\beta_0 = 0, \beta_1 = \bar{Y}$ ממזערים את הפונקציה G שהגדרנו), ולכן טיב הקשר מושלם, כלומר $R^2 = 1$. לעומת זאת אם $SSE = SST$, כלומר ישר הריבועים הפחותים הוא הישר הקבוע $y = \bar{Y}$, אזי טיב הקשר הוא הגרוע ביותר, כלומר $R^2 = 0$.

2.2 תכונות ישר הריבועים הפחותים

טענה: מצאנו כי $\beta_1^* = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$. ניתן לבטא ערך זה גם בשתי הצורות הבאות:

$$\beta_1^* = \frac{s_Y}{s_X} \cdot r_{X,Y}$$

$$\beta_1^* = \frac{nS_{XY} - S_X S_Y}{nS_{XX} - S_X^2}$$

(ההוכחה נובעת מחישוב ישיר של הביטויים, והיא מושארת כתרגיל).

טענה: יהי $(X_1, Y_1), \dots, (X_n, Y_n)$ מדגם, ויהיו (β_0^*, β_1^*) המגדירים את ישר הריבועים הפחותים המתאים למדגם. אזי לכל $c > 0$ מתקיים:

• עבור המדגם $(X_1, cY_1), \dots, (X_n, cY_n)$, ישר הריבועים הפחותים מתקבל על ידי $(c\beta_0^*, c\beta_1^*)$.

- עבור המדגם $(cX_1, Y_1), \dots, (cX_n, Y_n)$, ישר הריבועים הפחותים מתקבל על ידי $(\beta_0^*, \frac{1}{c}\beta_1^*)$.

הוכחה: לצורך טענה זו נוח להתבונן בצורה $\beta_1^* = \frac{s_Y}{s_X} \cdot r_{X,Y}$. נזכור כי $r_{X,Y}$ אינו משתנה בעקבות טרנספורמציה לינארית ששומרת סימן. נתבונן בשני חלקי הטענה:

- במקרה זה ברור כי s_X לא משתנה, וכן כי $s_{cY} = cs_Y$. מכאן כי $\beta_1^* \mapsto \frac{cs_Y}{cs_X} r_{X,Y} = c\beta_1^*$.
כמו כן ברור כי $\bar{X}$ לא משתנה, וכן כי $c\bar{Y} = \overline{cY}$, ולכן מהנוסחה $\beta_0^* = \bar{Y} - \beta_1^* \bar{X}$ נובע כי $\beta_0^* \mapsto c\bar{Y} - c\beta_1^* \bar{X} = c\beta_0^*$.
- במקרה זה ברור כי s_Y לא משתנה, וכן כי $s_{cX} = cs_X$. מכאן כי $\beta_1^* \mapsto \frac{s_Y}{cs_X} \cdot r_{X,Y} = \frac{1}{c}\beta_1^*$.
כמו כן ברור כי $\bar{Y}$ לא משתנה, וכן כי $c\bar{X} = \overline{cX}$, ולכן מהנוסחה $\beta_0^* = \bar{Y} - \beta_1^* \bar{X}$ נובע כי $\beta_0^* \mapsto \bar{Y} - \frac{1}{c}\beta_1^* \cdot c\bar{X} = \bar{Y} - \beta_1^* \bar{X} = \beta_0^*$. ■

תכונות נוספות: להלן כמה תכונות פשוטות נוספות של ישר הריבועים הפחותים:

$$1. \sum_{i=1}^n e_i = 0. \text{ כלומר סכום השגיאות הוא אפס.}$$

הוכחה: נחשב:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n (Y_i - \beta_0^* - \beta_1^* X_i) = 0$$

כאשר השוויון האחרון נובע מכך שבחירת β_0^*, β_1^* איפסה את הנגזרת של הפונקציה $G(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i)^2$, וקל לראות שהנגזרת של פונקציה זו שווה לביטוי הנ"ל כפול קבוע כלשהו.

$$2. \text{ הנקודה } (\bar{X}, \bar{Y}) \text{ נמצאת על ישר הריבועים הפחותים.}$$

הוכחה: צריך להראות כי $\bar{Y} = \beta_0^* + \beta_1^* \bar{X}$. אבל נזכור כי $\beta_0^* = \bar{Y} - \beta_1^* \bar{X}$, ולכן השוויון ברור.

$$3. \bar{\hat{Y}} = \bar{Y}. \text{ כלומר הממוצע של הערכים החזויים שווה לממוצע של הערכים הנצפים.}$$

הוכחה: נחשב:

$$\bar{\hat{Y}} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \frac{1}{n} \sum_{i=1}^n (\beta_0^* + \beta_1^* X_i) = \beta_0^* + \beta_1^* \bar{X} = \bar{Y}$$

3 רגרסיה לינארית פשוטה - מודל סטטיסטי כללי לשגיאות

מבוא: עד עכשיו הגדרנו סטטיסטיים שנובעים מהמדגם $(X_1, Y_1), \dots, (Y_1, Y_n)$, אולם לא הנחנו שום דבר סטטיסטי, כמו למשל ההתפלגות שממנה מגיע המדגם, התוחלת וכדומה. נניח כעת מודל שבו מתקיימות הנחות מסוימות לגבי המדגם.

המודל: בהינתן מדגם $(X_1, Y_1), \dots, (X_n, Y_n)$ של מספרים ממשיים, נתייחס אל $X_1, \dots, X_n$ כאל מספרים ונתייחס אל $Y_1, \dots, Y_n$ כאל משתנים מקריים.

נניח שקיימים β_0, β_1 כלשהם, כך שלכל $i = 1, \dots, n$ מתקיים $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ כאשר ϵ_i הוא משתנה מקרי כלשהו, המקיים את שלוש התכונות הבאות:

$$1. E[\epsilon_i] = 0 \text{ לכל } i$$

$$2. Cov(\epsilon_i, \epsilon_j) = 0 \text{ לכל } i \neq j$$

$$3. Var(\epsilon_i) = Var(\epsilon_j) \text{ לכל } i \neq j$$

היות וכל השונות של $\epsilon_1, \dots, \epsilon_n$ שוות, נסמן את השונות של כולם על ידי σ^2 .

נשים לב כי היות והנחנו $E[\epsilon_i] = 0$, נובע כי $E[\epsilon_i^2] = Var(\epsilon_i) = \sigma^2$.

סימון: כעת משיש לנו מודל סטטיסטי, נתייחס אל β_0^*, β_1^* שמגדירים את ישר הריבועים הפחותים כאל אומדים של β_0, β_1 , ונסמן אותם מחדש $\hat{\beta}_0 := \beta_0^*, \hat{\beta}_1 := \beta_1^*$.

הגדרה: עבור $n > 2$, נגדיר את הביטוי $s^2 := \frac{SSE}{n-2}$ להיות אומדן עבור השונות σ^2 .

הערה: נזכור שהגדרנו את השגיאה $e_i = Y_i - \hat{Y}_i$ כאשר $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. נשים לב שבמסגרת המודל ניתן לחשוב על ϵ_i כעל אומדן של e_i .

תזכורת: באופן כללי, עבור משתנים מקריים $Z_1, \dots, Z_n$, מגדירים את מטריצת השונות A על ידי $A_{ij} := Cov(Z_i, Z_j)$. כלומר, האיבר בשורה ה- i ובעמודה ה- j הוא $Cov(Z_i, Z_j)$.

נשים לב שאיברי האלכסון של המטריצה, היכן שמתקיים $i = j$, הם $Cov(Z_i, Z_i) = Var(Z_i)$.

סימון: נסמן בכתוב ווקטורי $\vec{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$, ונסמן את מטריצת היחידה $n \times n$ על ידי $I_{n \times n}$.

נסמן את מטריצת השונות של $\vec{\epsilon}$ על ידי $Cov(\vec{\epsilon})$, ונשים לב שמהנחת המודל נובע $Cov(\vec{\epsilon}) = \sigma^2 I_{n \times n}$.

כלומר, $Cov(\vec{\epsilon})$ היא אפס בכל מקום מחוץ לאלכסון (כי הנחנו $Cov(\epsilon_i, \epsilon_j) = 0$ לכל $i \neq j$), והיא σ^2 על האלכסון (כי הנחנו $Var(\epsilon_i) = \sigma^2$ לכל i).

3.1 תוחלות של האומדים

טענה: האומדים $\hat{\beta}_0, \hat{\beta}_1, s^2$ חסרי הטיות: $E[\hat{\beta}_0] = \beta_0, E[\hat{\beta}_1] = \beta_1, E[s^2] = \sigma^2$.

הוכחה: את חוסר ההטיה של האומדן לשונות נוכיח בהמשך הקורס. נוכיח את חוסר ההטיה של שני האומדים האחרים.

1. נראה כי $E[\hat{\beta}_1] = \beta_1$. תחילה נשים לב שהיות והנחת המודל היא כי X_i כולם מספרים קבועים וכי Y_i משתנים מקריים, נובע כי:

$$E[\hat{\beta}_1] = E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] = \frac{\sum_{i=1}^n (X_i - \bar{X}) E[Y_i - \bar{Y}]}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

כעת נשתמש בהנחת המודל $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, ונחשב את הביטוי הבא:

$$\begin{aligned} E[Y_i - \bar{Y}] &= E\left[Y_i - \frac{1}{n} \sum_{j=1}^n Y_j\right] = \\ &= E\left[\beta_0 + \beta_1 X_i + \epsilon_i - \frac{1}{n} \sum_{j=1}^n (\beta_0 + \beta_1 X_j + \epsilon_j)\right] = \\ &= E[\beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 - \beta_1 \bar{X} - \bar{\epsilon}] = \\ &= E[\beta_1 X_i - \beta_1 \bar{X}] = \beta_1 (X_i - \bar{X}) \end{aligned}$$

כאשר השוויון שלפני האחרון נובע מהנחת המודל כי $E[\epsilon_i] = 0$, והשוויון האחרון מכך שכל X_i הם מספרים קבועים. אם כך נציב את הערך שמצאנו בשוויון לעיל, ונקבל כי:

$$E[\hat{\beta}_1] = \frac{\sum_{i=1}^n (X_i - \bar{X}) \beta_1 (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1$$

כלומר, $\hat{\beta}_1$ אומד חסר הטיות של β_1 .

2. נראה כי $E[\hat{\beta}_0] = \beta_0$. נחשב:

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{X}] = E[\bar{Y}] - E[\hat{\beta}_1] \cdot \bar{X} = E[\bar{Y}] - \beta_1 \bar{X} = \\ &= \frac{1}{n} \sum_{i=1}^n E[Y_i] - \beta_1 \bar{X} = \frac{1}{n} \sum_{i=1}^n E[\beta_0 + \beta_1 X_i + \epsilon_i] - \beta_1 \bar{X} = \\ &= \beta_0 + \beta_1 \cdot \frac{1}{n} \sum_{i=1}^n X_i + \sum_{i=1}^n E[\epsilon_i] - \beta_1 \bar{X} = \beta_0 \end{aligned}$$

כאשר השוויון השני נובע מהנחת המודל כי X_i מספרים קבועים, השוויון השלישי נובע מכך שראינו כי $E[\hat{\beta}_1] = \beta_1$, והשוויון האחרון נובע מהנחת המודל כי $E[\epsilon_i] = 0$. ■

3.2 שונות של האומדים

טענה: $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$

הוכחה: הגדרנו $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$, ונשים לב כי ניתן לכתוב זהותית:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \frac{\overbrace{\sum_{i=1}^n (X_i - \bar{X})}^{=0}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

כאשר השוויון השני נובע מכך שהמחומר השני הוא 0.

אם כך נחשב את השונות המבוקשת:

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \frac{1}{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^2} Var\left(\sum_{i=1}^n (X_i - \bar{X}) Y_i\right)$$

כאשר השוויון השני נובע מכך שבמסגרת המודל מתייחסים אל X_i כאל קבועים.

תזכורת: באופן כללי עבור מ"מ $Z_1, \dots, Z_k$ מתקיימת הנוסחה:

$$Var\left(\sum_{i=1}^k Z_i\right) = \sum_{i=1}^k Var(Z_i) + \sum_{i \neq j} Cov(Z_i, Z_j)$$

כעת נסיק:

$$\begin{aligned} Var(\hat{\beta}_1) &= \\ &= \frac{1}{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^2} \sum_{i=1}^n Var(X_i - \bar{X}) Y_i + \sum_{i \neq j} (X_i - \bar{X})(X_j - \bar{X}) Cov(Y_i, Y_j) = \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^2} Var(Y_i) + \sum_{i \neq j} (X_i - \bar{X})(X_j - \bar{X}) Cov(Y_i, Y_j) = \\ &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} Var(Y_i) + \sum_{i \neq j} (X_i - \bar{X})(X_j - \bar{X}) Cov(Y_i, Y_j) \end{aligned}$$

כאשר השוויון הראשון הוא הפעלת הנוסחה שבתזכורת, השוויון השני הוא שוב הוצאת X_i כקבועים, והשוויון האחרון ברור.

כדי להשלים את הטענה נותר להראות כי $Var(Y_i) = \sigma^2$ וכי $Cov(Y_i, Y_j) = 0$.
נראה את שני השוויונים:

1. נשים לב כי $\beta_0 + \beta_1 X_i$ הוא מספר כלשהו, ולכן הוא לא משנה את השונות.
מכאן כי:

$$Var(Y_i) = Var(\beta_0 + \beta_1 X_i + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

2. מאותו נימוק נובע גם כי:

$$Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 X_i + \epsilon_i, \beta_0 + \beta_1 X_j + \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$$

■ וקל לראות כי שני שוויונים אלה משלימים את הטענה.

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \quad \text{טענה:}$$

הוכחה: הגדרנו $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, לכן נובע:

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{Y} - \hat{\beta}_1 \bar{X}) = Var(\bar{Y}) + Var(\hat{\beta}_1 \bar{X}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{X}) = \\ &= Var\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) + \bar{X}^2 Var(\hat{\beta}_1) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{X}) = \\ &= \frac{1}{n^2} \sum_{i=1}^n Var(Y_i) + \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} - 2\bar{X} Cov(\bar{Y}, \hat{\beta}_1) = \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] - 2\bar{X} Cov(\bar{Y}, \hat{\beta}_1) \end{aligned}$$

נותר להראות כי $Cov(\bar{Y}, \hat{\beta}_1) = 0$, וזה ישלים את הטענה. נחשב:

$$\begin{aligned}
 Cov(\bar{Y}, \hat{\beta}_1) &= Cov(\beta_0 + \beta_1 \bar{X} + \bar{\epsilon}, \hat{\beta}_1) = Cov(\bar{\epsilon}, \hat{\beta}_1) = \\
 &= Cov\left(\bar{\epsilon}, \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = Cov\left(\bar{\epsilon}, \frac{\sum_{i=1}^n (X_i - \bar{X}) (\beta_0 + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \\
 &= Cov\left(\bar{\epsilon}, \frac{\sum_{i=1}^n (X_i - \bar{X}) \epsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} Cov\left(\bar{\epsilon}, \sum_{i=1}^n (X_i - \bar{X}) \epsilon_i\right) = \\
 &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} Cov\left(\frac{1}{n} \sum_{j=1}^n \epsilon_j, \sum_{i=1}^n (X_i - \bar{X}) \epsilon_i\right) = \\
 &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \frac{1}{n} \sum_{j=1}^n Cov\left(\epsilon_j, \sum_{i=1}^n (X_i - \bar{X}) \epsilon_i\right) = \\
 &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^n \underbrace{(X_i - \bar{X}) Cov(\epsilon_j, \epsilon_i)}_{=0} = 0
 \end{aligned}$$

■ ושוויון זה משלים את הטענה.

3.2.1 פירוק השונות לרכיבים

תזכורת: הגדרנו לעיל:

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

הגדרה:

$$SSR := \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

סטטיסטי זה מודד את אחוז השונות הכללית המוסבר על ידי ישר הריבועים הפחותים.

טענה: ניתן לפרק את השונות על ידי $SST = SSE + SSR$.

במילים: שונות המדגם שווה לשונות השארית ביחס לישר הריבועים הפחותים ועוד שונות התחזיות שעל ישר הריבועים הפחותים.

הוכחה: נחשב:

$$\begin{aligned} SST &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i) (\hat{Y}_i - \bar{Y}) = \\ &= SSE + SSR + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i) (\hat{Y}_i - \bar{Y}) \end{aligned}$$

נותר להראות כי המחובר האחרון הוא אפס. נחשב:

$$\begin{aligned} \sum_{i=1}^n (Y_i - \hat{Y}_i) (\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^n e_i (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) = \\ &= (\hat{\beta}_0 - \bar{Y}) \sum_{i=1}^n e_i + \hat{\beta}_1 \sum_{i=1}^n e_i X_i = 0 \end{aligned}$$

כאשר השוויון האחרון הוא מכך שמתקיים $\sum_{i=1}^n e_i = 0$, כפי שראינו בפרק על תכונות ישר הריבועים הפחותים, וכן $\sum_{i=1}^n e_i X_i = 0$, כפי שראינו בתרגיל. ■

הערה: מהטענה האחרונה נובע כי ניתן לכתוב:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SST - SSR}{SST} = \frac{SSR}{SST}$$

טענה: $r_{X,Y}^2 = R^2$

הוכחה: נזכור שהגדרנו את מקדם המתאם $r_{X,Y} := \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$

נתבונן במדגם חדש $(Y_1, \hat{Y}_1), \dots, (Y_n, \hat{Y}_n)$ ונתבונן במקדם המתאם של מדגם זה:

$$\begin{aligned} r_{Y, \hat{Y}} &= \frac{\sum_{i=1}^n (Y_i - \bar{Y}) (\hat{Y}_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2} \sqrt{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}} = \\ &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y} + e_i) (\hat{Y}_i - \bar{Y})}{\sqrt{SST \cdot SSR}} = \\ &= \frac{1}{\sqrt{SST \cdot SSR}} \left[\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i (\hat{Y}_i - \bar{Y}) \right] = \\ &= \frac{1}{\sqrt{SST \cdot SSR}} \left[SSR + \sum_{i=1}^n e_i \hat{Y}_i - \bar{Y} \cdot \sum_{i=1}^n e_i \right] = \frac{\sqrt{SSR}}{\sqrt{SST}} = \sqrt{R^2} \end{aligned}$$

כאשר השוויון הראשון נובע מכך שהראינו כי $\bar{\hat{Y}} = \bar{Y}$, השוויון השני נובע כי $Y_i = \hat{Y}_i + e_i$, השוויונים השלישי והרביעי ברורים, השוויון החמישי נובע כי $\sum_{i=1}^n e_i \hat{Y}_i = 0$, והשוויון האחרון נובע מהטענה הקודמת.

אם כך מצאנו כי $r_{Y, \hat{Y}}^2 = R^2$. לכן כדי להשלים את ההוכחה מספיק להראות כי $r_{Y, \hat{Y}} = \text{sign}(\hat{\beta}_1) \cdot r_{X, Y}$ (כאשר $\text{sign}(\hat{\beta}_1)$ הוא 1 אם $\hat{\beta}_1 > 0$ והוא -1 אם $\hat{\beta}_1 < 0$). שוויון זה מושאר כתרגיל. ■

4 גרסיה לינארית פשוטה - מודל סטטיסטי נורמלי לשגיאות

מבוא: הנחנו לעיל מודל סטטיסטי כללי לגבי המדגם, מבלי שהנחנו התפלגות מסוימת. במסגרת המודל הנורמלי נניח שפרמטרים מסוימים מפולגים נורמלית, כפי שנפרט.

המודל: בהינתן מדגם $(X_1, Y_1), \dots, (X_n, Y_n)$ של מספרים ממשיים, נתייחס אל $X_1, \dots, X_n$ כאל מספרים ונתייחס אל $Y_1, \dots, Y_n$ כאל משתנים מקריים.

נניח שקיימים β_0, β_1 כלשהם, כך שלכל $i = 1, \dots, n$ מתקיים $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ כאשר $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ וכן $\epsilon_1, \dots, \epsilon_n$ בלתי תלויים.

הערה: נשים לב שמודל זה אכן מקיים את כל שלוש ההנחות שדרשנו במודל הכללי: $E[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$ וכן היות והנחנו $\epsilon_1, \dots, \epsilon_n$ בלתי תלויים הם בפרט גם בלתי מתואמים, כלומר $Cov(\epsilon_i, \epsilon_j) = 0$ לכל $i \neq j$.

טענה: במסגרת המודל הנורמלי, גם $\hat{\beta}_0, \hat{\beta}_1$ הם מ"מ שמפולגים נורמלית.

מסקנה: היות וכבר מצאנו את הנוסחאות הכלליות לתוחלת והשונות של $\hat{\beta}_0, \hat{\beta}_1$, מכך שהם מפולגים נורמלית נוכל להסיק יותר מכך, שמתקיים:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]\right)$$

הוכחה: נשים לב כי טרנספורמציה לינארית של מ"מ נורמלי מהווה מ"מ נורמלי. אם כך, היות ומתקיים $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ וכן במסגרת המודל הביטוי $\beta_0 + \beta_1 X_i$ הוא קבוע, נובע כי $Y_i \sim \mathcal{N}(\cdot, \cdot)$.

מכך נובע כי $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} Y_i \sim \mathcal{N}(\cdot, \cdot)$ כי זה כפל בקבוע של מ"מ נורמלי, וכעת נוכל גם להסיק כי $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \sim \mathcal{N}(\cdot, \cdot)$. ■

4.1 אומדי נראות מקסימלית

תזכורת: בהינתן מדגם $Z_1, \dots, Z_k$ של מ"מ ב"ת, כאשר $Z_i \sim f_i(z; \theta)$ (עבור f_i פונקציית צפיפות), מגדירים את **פונקציית הנראות** להיות:

$$L(z_1, \dots, z_k; \theta) := \prod_{i=1}^k f_i(z_i, \theta)$$

ומסמנים את לוג פונקציית הנראות:

$$l(z_1, \dots, z_k; \theta) := \log L(z_1, \dots, z_k; \theta)$$

במקרה כזה, מגדירים **אומד נראות מקסימלית** לפרמטר θ להיות:

$$\hat{\theta} := \arg \max_{\theta} l(z_1, \dots, z_k; \theta)$$

טענה: במסגרת המודל הנורמלי, אומדי הריבועים הפחותים הם גם אומדי נראות מירבית. כלומר, $\hat{\beta}_0, \hat{\beta}_1$ שמצאנו שמזערים את פונקציית הריבועים הפחותים, הם ממקסמים את פונקציית הנראות.

הוכחה: נחשב עבור המדגם $(X_1, Y_1), \dots, (X_n, Y_n)$ את פונקציית הנראות:

$$L((X_1, Y_1), \dots, (X_n, Y_n); \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f_i(X_i, Y_i; \beta_0, \beta_1, \sigma^2) =$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

כאשר השוויון הראשון נובע מכך שמניחים כי המ"מ במדגם ב"ת ולכן הצפיפות המשותפת היא מכפלת הצפיפויות השוליות, והשוויון השני נובע מהנחת הנורמליות, שכן כפי שהראינו כי $Y_i \sim \mathcal{N}$. נוציא $\log$ ונקבל:

$$l((X_1, Y_1), \dots, (X_n, Y_n); \beta_0, \beta_1, \sigma^2) =$$

$$= \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 =$$

$$= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2} G(\beta_0, \beta_1) =$$

כאשר $G(\beta_0, \beta_1) = \sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i)^2$ היא פונקציית סכום הריבועים (אז חילקנו ב- n , אולם לצורך מזעור ומקסום אין הבדל), אותה מיזערנו כדי למצוא את אומדי הריבועים הפחותים $\hat{\beta}_0, \hat{\beta}_1$.

לכן מהשוויון האחרון נובע כי היות והאומדים $\hat{\beta}_0, \hat{\beta}_1$ ממזערים את G , הם בהכרח ממקסמים את l (שכן l תלויה בהם בסימן שלילי). אבל אומדי נראות מקסימלית מוגדרים להיות אלה שממקסמים את l , ולכן תחת הנחת הנורמליות נובע כי $\hat{\beta}_0, \hat{\beta}_1$ הם גם אומדי נראות מקסימלית. ■

4.2 הסקה סטטיסטית

4.2.1 רווח סמך של β_1

מבוא: הראינו שבמודל הנורמלי מתקיים כי $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{s_X}\right)$ (כאשר $s_X := \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$). מכאן נובע שניתן לנרמל ולקבל כי $\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{n} \cdot s_X} \sim \mathcal{N}(0, 1)$. אולם הערך של σ אינו ידוע, ולכן נציב במקומו את האומד s .

טענה 1: עבור האומד $s^2 := \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ מתקיים כי $s^2 \sim \frac{n-2}{\sigma^2} \cdot \mathcal{X}_{(n-2)}^2$.

(נוכיח טענה זו בהמשך באופן כללי יותר).

טענה 2: יהיו Z, U מ"מ בלתי תלויים כך שמתקיים $Z \sim \mathcal{N}(0, 1)$ וכן $U \sim \mathcal{X}_{(m)}^2$, אזי $\frac{Z}{\sqrt{U/m}} \sim t_{(m)}$, כאשר $t_{(m)}$ היא התפלגות t -student עם m דרגות חופש. (ראינו טענה זו בקורסים קודמים).

טענה 3: האומדים $\hat{\beta}_1, s^2$ הם בלתי תלויים. (לא נוכיח טענה זו).

מסקנה: מתקיים כי $\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{n \cdot s_x}} \sim t_{(n-2)}$

הוכחה: כפי שהזכרנו במבוא מתקיים $\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{n \cdot s_x}} \sim \mathcal{N}(0, 1)$. כמו כן מטענה 1 מתקיים $s^2 \sim \mathcal{X}_{(n-2)}^2 \cdot \frac{n-2}{\sigma^2}$. לכן מטענה 2 (שמטענה 3 נובע שניתן להשתמש בה במקרה שלנו) נובע כי:

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{n \cdot s_x}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{n \cdot s_x}} / \sqrt{\frac{n-2}{n-2} \cdot \frac{s^2}{\sigma^2}} \sim t_{(n-2)}$$

כאשר השוויון נובע מצמצום. ■

סימון: נניח כי $T \sim t_{(m)}$ כאשר $t_{(m)}$ היא התפלגות t -student בעלת m דרגות חופש. נסמן את פונקציית ההתפלגות המצטברת שלו על ידי $F_T(x) = P(T \leq x)$. בהמשך לכך, בהינתן $0 < \alpha < 1$, נסמן על ידי $t_{m,\alpha}$ את הערך המקיים $F_T(t_{m,\alpha}) = \alpha$.

מסקנה: מתקבל רווח סמך ברמת סמך $1 - \alpha$ עבור β_1 , על ידי $\left[\hat{\beta}_1 \pm \frac{s}{\sqrt{n \cdot s_x}} \cdot t_{n-2, 1-\frac{\alpha}{2}} \right]$. **הוכחה:** צריך להראות כי ההסתברות שהפרמטר β_1 שייך לרווח הסמך היא $1 - \alpha$. כלומר:

$$P\left(\hat{\beta}_1 - \frac{s}{\sqrt{n \cdot s_x}} \cdot t_{n-2, 1-\frac{\alpha}{2}} \leq \beta_1 \leq \hat{\beta}_1 + \frac{s}{\sqrt{n \cdot s_x}} \cdot t_{n-2, 1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

נסמן $T = \frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{n \cdot s_x}}}$, ובאמצעות העברת אגפים נקבל שהסתברות זו היא:

$$\begin{aligned} P\left(-\frac{s}{\sqrt{n \cdot s_x}} \cdot t_{n-2, 1-\frac{\alpha}{2}} \leq \hat{\beta}_1 - \beta_1 \leq \frac{s}{\sqrt{n \cdot s_x}} \cdot t_{n-2, 1-\frac{\alpha}{2}}\right) &= \\ &= P\left(-t_{n-2, 1-\frac{\alpha}{2}} \leq T \leq t_{n-2, 1-\frac{\alpha}{2}}\right) = \\ &= F_T\left(t_{n-2, 1-\frac{\alpha}{2}}\right) - F_T\left(-t_{n-2, 1-\frac{\alpha}{2}}\right) = \\ &= F_T\left(t_{n-2, 1-\frac{\alpha}{2}}\right) - (1 - F_T\left(t_{n-2, 1-\frac{\alpha}{2}}\right)) = 2 \cdot F_T\left(t_{n-2, 1-\frac{\alpha}{2}}\right) - 1 = \\ &= 2 \cdot \left(1 - \frac{\alpha}{2}\right) - 1 = 2 - \alpha - 1 = 1 - \alpha \end{aligned}$$

כאשר השוויון הראשון הוא העברת אגפים, השוויון השני הוא סימון בלבד, השוויון השלישי נובע מתכונות של פונקציות התפלגות מצטברת, השוויון הרביעי נובע מסימטריות של התפלגות t_{n-2} , השוויון החמישי נובע מהגדרת $t_{n-2, 1-\frac{\alpha}{2}}$, ושאר השוויונים ברורים. ■

4.2.2 בדיקת השערות על β_1

עבור בדיקת ההשערות $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$, מתקבלת רמת מובהקות α תחת המדיניות של דחייה אם $|\hat{\beta}_1| > \frac{s}{\sqrt{n} \cdot s_x} \cdot t_{n-2, 1-\frac{\alpha}{2}}$

חלק II

רגרסיה לינארית מרובת משתנים

5 מבוא

ברגרסיה לינארית פשוטה הנחנו כי Y מוסבר על ידי משתנה אחד X בצורה $Y = \beta_0 + \beta_1 X + \epsilon$, כאשר β_0, β_1 מקדמים ממשיים כלשהם, וכן ϵ שגיאה. לעומת זאת ברגרסיה לינארית מרובת משתנים מניחים כי Y מוסבר על ידי p משתנים בצורה: $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

כאשר $\beta_0, \dots, \beta_p$ מקדמים ממשיים כלשהם, וכן ϵ שגיאה. באמצעות מדגם נתון נהיה מעוניינים לאמוד את המקדמים $\beta_0, \dots, \beta_p$.

הערה: בפרק הקרוב נשנה את הסימונים והאינדקסים של β_i לצורך הייצוג המטריציוני, ולכן הסימון שבמבוא זה הוא להסבר בלבד.

5.1 רגרסיה לינארית מרובת משתנים - הצגה מטריציונית ומודל סטטיסטי נורמלי

נניח כי נתון מודל לינארי מדרגה $p-1$, כלומר:

$$Y = \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \beta_p + \epsilon$$

(לצורך הנוחות סימנו את החותך להיות β_p , נראה בהמשך מדוע). נניח כי בידנו מדגם מגודל n המגיע מהמודל, מהצורה $(\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n)$, כאשר $Y_i \in \mathbb{R}$ וכן $\vec{X}_i \in \mathbb{R}^{p-1}$ עבור $i = 1, \dots, n$. כלומר $\vec{X}_i = (X_{i1}, \dots, X_{i,p-1})$. מהנחת המודל הלינארי נובע שלכל $i = 1, \dots, n$ מתקיים:

$$Y_i = X_{i1}\beta_1 + \dots + X_{i,p-1}\beta_{p-1} + \beta_p + \epsilon_i = \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i$$

לצורך הנוחות (מיד נראה מדוע), נוסיף לווקטורים הללו עוד רכיב קבוע 1, כלומר נגדיר מחדש $\vec{X}_i \in \mathbb{R}^p$, ונשים לב שעתה $\vec{X}_i = (X_{i1}, \dots, X_{i,p-1}, 1)$.

סימון: נסמן את המטריצה:

$$\mathbb{X} := \begin{bmatrix} \vec{X}_1 \\ \vdots \\ \vec{X}_n \end{bmatrix} = \begin{bmatrix} X_{11} & \dots & X_{1,p-1} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{n1} & \dots & X_{n,p-1} & 1 \end{bmatrix}_{n \times p}$$

כמו כן נסמן את הווקטורים:

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \vec{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

בכתיב מטריציוני נוכל לכתוב מחדש את המודל:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^p \beta_j X_{1j} + \epsilon_1 \\ \vdots \\ \sum_{j=1}^p \beta_j X_{nj} + \epsilon_n \end{bmatrix} = \begin{bmatrix} X_{11} & \dots & X_{1,p-1} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{n1} & \dots & X_{n,p-1} & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

ובקיצור, המודל הלינארי הוא:

$$\vec{Y} = \mathbb{X}\vec{\beta} + \vec{\epsilon}$$

הערה: נשים לב כי עבור $p = 2$ מתקבלת גרסיה לינארית פשוטה.

מודל נורמלי לשגיאות: כאן נתחיל ישר לטפל במודל נורמלי לשגיאות, כלומר נניח כי $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ מ"מ בלתי תלויים.

בכתיב מטריציוני הנחה זו היא $\vec{\epsilon} \sim \mathcal{N}(\vec{0}, \sigma^2 I_{n \times n})$, כאשר:

$$\vec{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1} \quad \sigma^2 I_{n \times n} = \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix}_{n \times n}$$

6 שיטת הריבועים הפחותים

נעדכן את הסימונים באופן שיהיה נוח יותר לחישוב. נציג את המדגם על ידי $Y_i = \sum_{j=1}^p X_{ij}\beta_j + \epsilon_j$, ובאופן מטריציאלי על ידי $Y = \mathbb{X}\beta + \epsilon$. נניח את המודל $Y = \sum_{j=1}^p X_j \beta_j$, ונרצה למצוא אומד לוקטור המקדמים β . לעיל מצאנו אומד ריבועים פחותים $\hat{\beta}$ עבור $p = 1$. נרצה למצוא אומד ריבועים פחותים במקרה הכללי. נזכור כי אומד ריבועים פחותים הוא הערך שממזער את פונקציית השגיאות הריבועיות. אם כך נסמן את הפונקציה הזו:

$$F(\beta) = \sum_{i=1}^n \left(\sum_{j=1}^p X_{ij}\beta_j - Y_i \right)^2$$

ובהתאם נגדיר אומד $\hat{\beta}$ לוקטור β , על ידי:

$$\hat{\beta} = \arg \min_{\beta} F(\beta)$$

משפט: אם המטריצה $\mathbb{X}$ היא מדרגה מלאה, כלומר $rank(\mathbb{X}) = p$, אזי $\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y$.

למשפט זה נראה שתי הוכחות שונות. הראשונה תהיה ישירה וטכנית באמצעות גזירת הפונקציה $F(\beta)$, והשנייה תהיה באמצעות הטלות, כפי שנסביר בהמשך.

³דרגה של מטריצה היא ממד המרחב שנפרש על ידי עמודותיה.

6.1 הוכחה ראשונה: באמצעות גזירה

6.1.1 מבוא: מושגים כלליים באינפי של כמה משתנים

בהוכחה זו נעשה שימוש במשפט אודות נגזרות של פונקציות בכמה משתנים, ולשם כך נזכיר מושגים יסודיים.

הגדרות: תהי $g : \mathbb{R}^k \rightarrow \mathbb{R}$ פונקציה גזירה פעמיים. נסמן משתנה כללי $x = (x_1, \dots, x_k) \in \mathbb{R}^k$.

• נגדיר את הגרדיאנט של g להיות וקטור הנגזרות החלקיות שלה, כלומר:

$$\nabla g(z) = \left(\frac{\partial}{\partial x_1} g(z), \dots, \frac{\partial}{\partial x_k} g(z) \right)$$

כאשר $\frac{\partial}{\partial x_j} g$ היא הנגזרת החלקית ה- j , כלומר הנגזרת של g לפי המשתנה x_j .

• נגדיר את ההסיאן של g להיות מטריצת הנגזרות החלקיות השניות שלה, כלומר:

$$\nabla^2 g(z) = \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} g(z) & \dots & \frac{\partial^2}{\partial x_1 \partial x_k} g(z) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_k \partial x_1} g(z) & \dots & \frac{\partial^2}{\partial x_k \partial x_k} g(z) \end{pmatrix}$$

כאשר $\frac{\partial^2}{\partial x_j \partial x_k}$ היא הנגזרת החלקית ה- k של הנגזרת החלקית ה- j . כלומר הנגזרת של $\frac{\partial}{\partial x_j} g(z)$ לפי המשתנה x_k .

הגדרה: תהי $A_{k \times k}$ מטריצה ריבועית כלשהי. אומרים כי A מוגדרת חיובית אם לכל $z \in \mathbb{R}^k$ מתקיים $z^T A z \geq 0$. במקרה כזה מסמנים $A \succ 0$.

באופן סימטרי, אומרים כי A מוגדרת שלילית אם לכל $z \in \mathbb{R}^k$ מתקיים $z^T A z \leq 0$. במקרה כזה מסמנים $A \prec 0$.

משפט: תהי $g : \mathbb{R}^k \rightarrow \mathbb{R}$ פונקציה גזירה פעמיים. תהי $z^0 \in \mathbb{R}^k$ שעבורה מתקיים $\nabla g(z^0) = 0$.

אם מתקיים $\nabla^2 g(z^0) \succ 0$ אז z^0 היא נקודת מינימום מקומית.

אם מתקיים $\nabla^2 g(z^0) \prec 0$ אז z^0 היא נקודת מינימום מקומית.

6.1.2 הוכחת משפט הריבועים הפחותים

נשים לב כי הפונקציה $F(\beta) = \sum_{i=1}^n \left(\sum_{j=1}^p X_{ij} \beta_j - Y_i \right)^2$ היא מהצורה $F : \mathbb{R}^p \rightarrow \mathbb{R}$. אם כך נסמן משתנה כללי שלה על ידי $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$, ונחשב את הנגזרת ה- j .

שלה:

$$\begin{aligned}\frac{\partial}{\partial \beta_j} F(\beta) &= \frac{\partial}{\partial \beta_j} \sum_{i=1}^n \left(\sum_{j=1}^p X_{ij} \beta_j - Y_i \right)^2 = \sum_{i=1}^n \frac{\partial}{\partial \beta_j} \left(\sum_{j=1}^p X_{ij} \beta_j - Y_i \right)^2 = \\ &= \sum_{i=1}^n 2 \left(\sum_{j=1}^p X_{ij} \beta_j - Y_i \right) \cdot X_{ij} = 2 \sum_{i=1}^n \sum_{j=1}^p X_{ij} X_{ij} \beta_j - 2 \sum_{i=1}^n X_{ij} Y_i = \\ &= 2 [\mathbb{X}^T \mathbb{X} \beta]_j - 2 [\mathbb{X}^T Y]_j = 2 ([\mathbb{X}^T \mathbb{X} \beta]_j - [\mathbb{X}^T Y]_j)\end{aligned}$$

זה נכון עבור $j = 1, \dots, p$, ולכן הגרדיאנט הוא:

$$\nabla F(\beta) = 2 ([\mathbb{X}^T \mathbb{X} \beta]_1 - [\mathbb{X}^T Y]_1, \dots, [\mathbb{X}^T \mathbb{X} \beta]_p - [\mathbb{X}^T Y]_p) = 2 [\mathbb{X}^T \mathbb{X} \beta - \mathbb{X}^T Y]$$

נשווה לאפס, ונקבל:

$$0 = \nabla F(\hat{\beta}) = 2 [\mathbb{X}^T \mathbb{X} \hat{\beta} - \mathbb{X}^T Y]$$

$\Downarrow$

$$\mathbb{X}^T \mathbb{X} \hat{\beta} = \mathbb{X}^T Y$$

מהנחה כי $rank(\mathbb{X}) = p$ נובע כי $\mathbb{X}^T \mathbb{X}$ מטריצה הפיכה, ולכן נוכל להכפיל במטריצה $(\mathbb{X}^T \mathbb{X})^{-1}$ את שני האגפים ולקבל את הפתרון:

$$\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y$$

אם כך $\hat{\beta}$ שבמשפט היא אכן נקודת קיצון של $F(\beta)$. נותר לוודא שזו נקודת מינימום. תחילה נחשב את הנגזרת החלקית השנייה ה- k של $F(\beta)$. כלומר נגזור את הנגזרת ה- j לפי המשתנה ה- k :

$$\frac{\partial^2}{\partial \beta_k \partial \beta_j} F(\beta) = \frac{\partial}{\partial \beta_k} \left(2 \sum_{i=1}^n \sum_{j=1}^p X_{ij} X_{ij} \beta_j - 2 \sum_{i=1}^n X_{ij} Y_i \right) = 2 X_{ik} X_{ik}$$

$\Downarrow$

$$\nabla^2 F(\beta) = 2 \mathbb{X}^T \mathbb{X}$$

כדי להראות כי $\hat{\beta}$ היא מינימום יש להראות שזו מטריצה מוגדרת חיובית.

למה: תהי $A_{n \times p}$ מטריצה מדרגה מלאה, כלומר $rank(A) = p$. אזי $A^T A$ היא מטריצה מוגדרת חיובית.

הוכחה: לכל $z \in \mathbb{R}^p$ מתקיים:

$$z^T A^T A z = (Az)^T Az = \|Az\|^2 \geq 0$$

■

אם כך נובע כי $\nabla^2 F(\beta) = 2\mathbb{X}^T \mathbb{X} \succ 0$, ולכן $\hat{\beta}$ היא נקודת מינימום מקומית. כדי להסיק כי היא נקודת מינימום גלובלית, נשים לב כי היא נקודת מינימום מקומית יחידה, וכן נשים לב כי $F(\hat{\beta}) \xrightarrow{\beta_j \rightarrow -\infty} -\infty$ וגם $F(\hat{\beta}) \xrightarrow{\beta_j \rightarrow \infty} \infty$ ■

6.2 הוכחה שנייה: באמצעות הטלות

6.2.1 מבוא: הטלות

הגדרה: יהיו $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$ וקטורים. נסמן $V := \text{Span}\{\vec{v}_1, \dots, \vec{v}_k\}$. יהי $u \in \mathbb{R}^d$ וקטור כלשהו. **ההטלה** של u על V היא הווקטור שהכי קרוב ל- u בתת המרחב V . כלומר, וקטור ההטלה הוא הווקטור $w \in V$ כך שהמרחק $\|u - w\|$ מינימלי ביחס לכל שאר הווקטורים ב- V .

הגדרה: תהי A מטריצה שעמודותיה בלתי תלויות לינארית.

נגדיר את **מטריצת ההטלה** של A להיות המטריצה:

$$P_A := A(A^T A)^{-1} A^T$$

טענה: תהי A מטריצה כמו בהגדרה האחרונה. אזי ההטלה של וקטור u על תת המרחב הנפרש על ידי העמודות של A , כלומר על $\text{Span}\{A\}$, הוא הווקטור $P_A u$.

תכונות: תהי A מטריצה שעמודותיה בלתי תלויות לינארית.

1. אם $u \in \text{Span}\{A\}$ אזי $P_A u = u$.

2. $P_A^2 = P_A$.

3. $P_A(I - P_A) = 0$.

4. $P_A^T(I - P_A) = 0$.

5. אם P_A מטריצת הטלה על $\text{Span}\{A\}$, אזי $I - P_A$ היא ההטלה על המרחב הניצב ל- $\text{Span}\{A\}$.

6.2.2 הוכחת משפט הריבועים הפחותים

מההגדרה מתקיים בצורה מטריציאלית:

$$\hat{\beta} = \arg \min_{\beta} F(\beta) = \arg \min_{\beta} \sum_{i=1}^n \left(\sum_{j=1}^p X_{ij} \beta_j - Y_i \right)^2 = \arg \min_{\beta} \|\mathbb{X}\beta - Y\|^2$$

מכאן נובע שעבור $\hat{Y} = \mathbb{X}\hat{\beta}$ מתקיים כי:

$$\hat{Y} = \mathbb{X}\hat{\beta} = \arg \min_{\mathbb{X}\beta} \|\mathbb{X}\beta - Y\|^2$$

נשים לב כי האוסף שעליו מבצעים את המינימוזציה הוא $\text{Span}(\mathbb{X})$, $\{\mathbb{X}\beta \mid \beta \in \mathbb{R}\}$, ולכן $\hat{Y}$ הוא למעשה ההטלה של Y על המרחב $\text{Span}(\mathbb{X})$.
אם כך נפעיל את מטריצת ההטלה $P_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ ונקבל כי:

$$\hat{Y} = \mathbb{X}\hat{\beta} = P_{\mathbb{X}}Y = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY$$

הנחנו כי $\mathbb{X}$ מטריצה מדרגה מלאה, נובע כי למערכת המשוואות $\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY$ קיים פתרון יחיד.

ברור כי $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY$ הוא פתרון, ומהיחידות נובע כי הוא פתרון יחיד. ■

6.3 הסקה סטטיסטית

6.3.1 התפלגות $\hat{\beta}$

תזכורת (המשפט הרב-נורמלי): יהי $W \sim \mathcal{N}(\mu, \Sigma)$ וקטור באורך d , תהי A מטריצה מגודל $m \times d$ ויהי c וקטור באורך m . אזי:

$$c + AW \sim \mathcal{N}(c + A\mu, A\Sigma A^T)$$

טענה: עבור $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY$ מתקיים $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$.

הוכחה: ידוע כי $Y \sim \mathcal{N}(\mathbb{X}\beta, \sigma^2 I)$. תחילה מתקיים כי $\hat{\beta} \sim \mathcal{N}(\cdot)$ שכן הוא טרנספורמציה ליניארית של מ"מ $y \sim \mathcal{N}(\cdot)$. כמו כן מתקיים:

$$E[\hat{\beta}] = E[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY] = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TE[Y] = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta = \beta$$

וכן מתקיים:

$$\begin{aligned} Var(\hat{\beta}) &= Var((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY) = \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TVar(Y)[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T]^T = \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\sigma^2\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} \end{aligned}$$

לכן נובע:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$$

■

6.3.2 אמידת σ^2

טענה: אומד של השונות σ^2 הוא $s^2 = \frac{1}{n-p} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 = \frac{1}{n-p} \|\hat{Y} - Y\|^2$
משפט: אם $\text{rank}(X) = p$ אז $\frac{n-p}{\sigma^2} s^2 \sim \chi^2_{(n-p)}$.

מסקנה: נובע כי רווח סמך של σ^2 ברמת סמך $1-\alpha$ מתקבל על ידי $\left[\frac{(n-p)s^2}{\chi^2_{n-p, 1-\alpha/2}}, \frac{(n-p)s^2}{\chi^2_{n-p, \alpha/2}} \right]$.

הוכחה: נגדיר את המטריצה $Q := I - P_X$, כאשר $P_X = X(X^T X)^{-1} X^T$ היא מטריצת ההטלה למרחב $\text{Span}(X)$. כלומר, Q היא מטריצת ההטלה למרחב הניצב $\text{Span}(X)^\perp$.

• חלק ראשון: תכונות המטריצה Q

נסמן את וקטור השאריות $e = Y - \hat{Y}$ ונסמן את וקטור השגיאות $\epsilon = Y - X\beta$.
 נראה כמה תכונות של המטריצה Q :

1. מתקיים $e = QY$, כי:

$$QY = (I - P_X)Y = Y - P_X Y = Y - X(X^T X)^{-1} X^T Y = Y - \hat{Y} = e$$

2. מתקיים $e = Q\epsilon$, כי:

$$e = QY = Q(X\beta + \epsilon) = QX\beta + Q\epsilon$$

כדי להשלים את השוויון המבוקש יש להראות כי $QX\beta = 0$, ואכן היות ומתקיים $P_X X = X$ מתכונות מטריצת ההטלה, נובע:

$$QX\beta = (I - P_X)X\beta = (X - P_X X)\beta = (X - X)\beta = 0$$

3. מתקיים $Q = Q^2$ וכן $Q = Q^T$, כפי שנובע מתכונות כלליות של מטריצות הטלה.

4. מתקיים $e^T e = \epsilon^T Q \epsilon$, כי:

$$e^T e = (Q\epsilon)^T Q\epsilon = \epsilon^T Q^T Q\epsilon = \epsilon^T Q\epsilon$$

5. מתקיים $\text{rank}(Q) = n - p$, כי כפי שנובע באופן כללי ממשפט הממדים:

$$\underbrace{\dim \text{Span}(X)}_{=\text{rank}(X)=p} + \dim \text{Span}(X)^\perp = \dim(\mathbb{R}^n) = n$$

$$\Downarrow$$

$$\text{rank}(Q) := \dim \text{Span}(X)^\perp = n - p$$

• חלק שני: ייצוג ספקטרלי של המטריצה Q

נשתמש במשפט הפירוק הספקטרלי. בהתאם למשפט זה, ניתן לפרק את המטריצה הסימטרית Q (כי $Q = Q^T$) לצורה $Q = U\Lambda U^T$, כאשר U היא מטריצה אורתוגונלית (כלומר $U^T U = I$) שהעמודות שלה הם הווקטורים העצמיים של Q , וכן Λ היא מטריצה שכולה אפסים חוץ מהאלכסון הראשי שלה, שמכיל את הערכים העצמיים של Q (כלומר היא מהצורה $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$). נראה שתי תכונות של הפירוק הספקטרלי של המטריצה Q :

1. **טענה:** כל הערכים העצמיים של המטריצה Q הם 0 או 1 בלבד.
הוכחה: יהי λ ערך עצמי כלשהו של Q . אז קיים וקטור עצמי מתאים u שעבורו $Qu = \lambda u$, ולכן:

$$\lambda u = Qu = Q^2 u = Q(Qu) = Q(\lambda u) = \lambda^2 u$$

לכן $\lambda = \lambda^2$, ולכן בהכרח $\lambda = 0$ או $\lambda = 1$. ■

2. **טענה:** קיימים בדיוק $n - p$ ערכים עצמיים של Q שהם 1.
הוכחה: מתקיים $\text{rank}(U\Lambda U^T) = \text{rank}(Q) = n - p$ אבל כפל במטריצה הפיכה לא משנה את הדרגה (ובפרט כל מטריצה אורתוגונלית היא הפיכה), ולכן נובע כי $\text{rank}(\Lambda) = n - p$. נשים לב כי היות והמטריצה Λ היא מטריצה שכולה אפסים למעט האלכסון הראשי שבו יש 0 או 1, בהכרח כי 1 מופיע בדיוק $n - p$ פעמים. ■

• חלק שלישי: ייצוג הביטוי $\frac{n-p}{\sigma^2} s^2$ נשים לב שמתקיים:

$$s^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-p} \sum_{i=1}^n e_i^2 = \frac{1}{n-p} e^T e$$

ולכן נובע:

$$(n-p) s^2 = e^T e$$

נסמן $\eta := U^T \epsilon$, ונשים לב שבהתאם לסימון זה מתקיים:

$$(n-p) s^2 = e^T e = \epsilon^T Q \epsilon = \epsilon^T U \Lambda U^T \epsilon = \eta^T \Lambda \eta$$

נזכור כי Λ היא מטריצה של אפסים בכל מקום למעט $n - p$ פעמים 1 באלכסון, ולכן נובע:

$$(n-p) s^2 = e^T e = \eta^T \Lambda \eta = \sum_{j=1}^{n-p} \eta_j^2$$

• חלק רביעי: התפלגות $\frac{n-p}{\sigma^2} s^2$

נזכור כי $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, ולכן מתכונות ההתפלגות הרב-נורמלית נובע כי:

$$\eta = U^T \epsilon \sim \mathcal{N}(U^T \cdot 0, U^T \sigma^2 (U^T)^T) = \mathcal{N}(0, U^T \sigma^2 U)$$

אבל נשים לב כי $U^T \sigma^2 U = \sigma^2 U^T U = \sigma^2 I$ ולכן:

$$\eta \sim \mathcal{N}(0, \sigma^2 I)$$

בפרט, לכל רכיב η_i בווקטור η מתקיים $\eta_i \sim \mathcal{N}(0, \sigma^2)$ ולכן נובע כי:

$$\frac{\eta_i}{\sigma^2} \sim \mathcal{N}(0, 1)$$

באופן כללי סכום של k ריבועי מ"מ נורמליים סטנדרטיים מתפלג $\chi^2_{(k)}$, ולכן מכל מה שראינו נובע:

$$\frac{(n-p)}{\sigma^2} s^2 = \frac{1}{\sigma^2} \sum_{j=1}^{n-p} \eta_j^2 = \sum_{j=1}^{n-p} \frac{\eta_j^2}{\sigma^2} \sim \chi^2_{(n-p)}$$

כפי שנדרש במשפט. ■

6.3.3 רווח סמך של רכיבי הווקטור β

ראינו שמתקיים $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$ ולכן לכל רכיב $\hat{\beta}_j$ בווקטור $\hat{\beta}$ (עבור $j = 1, \dots, p$) מתקיים $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 (\mathbb{X}^T \mathbb{X})_{jj}^{-1})$. לאחר תקנון נקבל:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{(\mathbb{X}^T \mathbb{X})_{jj}^{-1}}} \sim \mathcal{N}(0, 1)$$

אם נציב את האומד s^2 של σ^2 , נשים לב כי $s = \sqrt{\frac{1}{n-p} e^T e}$ ולכן נקבל:

$$\frac{\hat{\beta}_j - \beta_j}{s \sqrt{(\mathbb{X}^T \mathbb{X})_{jj}^{-1}}} \sim t_{(n-p)}$$

לכן מתקבל רווח סמך ברמת סמך $1-\alpha$ עבור $\hat{\beta}_j$, על ידי

$$\left[\hat{\beta}_j \pm s \sqrt{(\mathbb{X}^T \mathbb{X})_{jj}^{-1}} \cdot t_{n-p, 1-\alpha/2} \right]$$

6.3.4 בדיקת השערות על רכיבי הווקטור β

לכל רכיב $\hat{\beta}_j$ בווקטור $\hat{\beta}$ (עבור $j = 1, \dots, p$) עובר בדיקת ההשערות

$$\begin{cases} H_0 : \hat{\beta}_j = 0 \\ H_1 : \hat{\beta}_j \neq 0 \end{cases}$$

מתקבלת רמת מובהקות α תחת המדיניות של דחייה אם $|\hat{\beta}_j| > s \sqrt{(\mathbb{X}^T \mathbb{X})_{jj}^{-1}} \cdot t_{n-p, 1-\alpha/2}$

7 טרנספורמציות

בהינתן מדגם (X, Y) , מודל של רגרסיה לינארית רגילה קובע קשר בין Y לבין X , על ידי:

$$Y = \beta X + \epsilon$$

ואז באמצעות שיטת הריבועים הפחותים מקבלים אומד $\hat{\beta}$ שעבורו:

$$\hat{Y} = \hat{\beta}X$$

לעתים במקום זאת נרצה להניח מודל אחר של קשר בין Y לבין X , על ידי:

$$Y = \sum_{j=0}^m \beta_j h_j(X) + \epsilon$$

כאשר $h_0, \dots, h_m$ הן טרנספורמציות כלשהן (למשל $\log, \exp$) העלאה בחזקה, הוצאת שורש, או כל סוג אחר של טרנספורמציה). לאחר הפעלת הטרנספורמציות על המדגם המקורי (X, Y) , מתקבל מדגם חדש מהצורה $(h_0(X), \dots, h_m(X), Y)$, וכמו בגרסיה לינארית רגילה ניתן לקבל אומד $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_m)$ שעבורו:

$$\hat{Y} = \sum_{j=0}^m \hat{\beta}_j h_j(X)$$

דוגמה: (רגרסיה ריבועית) נבחר את הטרנספורמציות:

$$h_0(x) = 1$$

$$h_1(x) = x$$

$$h_2(x) = x^2$$

ונניח מודל ריבועי, כלומר:

$$Y = \sum_{j=0}^2 \beta_j h_j(X) + \epsilon = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

כעת באמצעות רגרסיה על המדגם $(1, X, X^2, Y)$ נקבל אומד $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ שעבורו:

$$\hat{Y} = \sum_{j=0}^2 \hat{\beta}_j h_j(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

הערה: לעתים נעדיף לבצע טרנספורמציות דווקא ל- Y . למשל $\log(Y) = \beta_0 + \beta_1 X + \epsilon$, ואז $Y = e^{\beta_0 + \beta_1 X} \cdot e^\epsilon$. נשים לב שבדוגמה זו מתקבל מודל שגיאה כפלי.

הערה: לעתים נרצה ליצור מודל עם אינטראקציה בין המשתנים. למשל $Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i}$.

מודל כזה לוקח בחשבון מצב בו השילוב של המשתנים X_{1i}, X_{2i} הוא בעל אפקט מובהק, בעוד שכל אחד מהמשתנים הללו עלול להיות חסר השפעה מובהקת.

8 משתנים קטגוריים

בהינתן מדגם (X, Y) , המשתנה X יכול להיות לא רציף אלא בעל מספר בדיד של קטגוריות.

המקרה הבינארי: נדון תחילה כאשר מדובר בשתי קטגוריות בלבד שנסמן A_1, A_2 . במקרה כזה נגדיר את המשתנה X שיקבל את הערך 0 אם מתקיימת אחת הקטגוריות ואת הערך 1 אם היא לא מתקיימת (ובמקרה הבינארי בהכרח זה אומר שמתקיימת הקטגוריה השנייה). כלומר $X := \mathbb{1}\{A_1\}$.⁴

כעת נוכל לבצע רגרסיה על (X, Y) , ובתוצאה של הרגרסיה נקבל אומד $\hat{\beta}$, שהמשמעות שלו היא השינוי בתוחלת של Y כאשר משנים את X מ-0 ל-1, כאשר כל שאר המשתנים קבועים וזהים.

דוגמה: A_1 הוא התכונה להיות זכר, A_2 הוא התכונה להיות נקבה. במקרה כזה נקודד את המאפיין "זכר" על ידי 0 ואת המשתנה "נקבה" על ידי 1 (למשל), ואז $X = \mathbb{1}\{A_2\}$ מקיים כי אם הנדגם זכר אז $X = 0$ ואם הנדגם נקבה אז $X = 1$.

המקרה הכללי: כעת נדון כאשר מדובר ב- k קטגוריות שנסמן $A_1, \dots, A_k$. במקרה כזה נגדיר $k-1$ משתנים חדשים $X_1, \dots, X_{k-1}$ שכל אחד מהם יקבל את הערך 0 אם מתקיימת הקטגוריה המתאימה לו ואת הערך 1 אם היא לא מתקיימת. כלומר $X_i := \mathbb{1}\{A_i\}$ עבור $i = 1, \dots, k-1$.

נשים לב שעבור הקטגוריה A_k לא הגדרנו משתנה מתאים, והיא מכונה **קטגוריית הרפרנס**.

כעת נוכל לבצע רגרסיה על $(X_1, \dots, X_{k-1}, Y)$, ובתוצאה של הרגרסיה נקבל אומד $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{k-1})$, שבו המשמעות של המקדם $\hat{\beta}_j$ היא השינוי בתוחלת של Y כאשר משנים את הקטגוריה A_j ביחס לקטגוריית הרפרנס A_k .

כלומר, אם אנחנו במצב $(0, \dots, 0)_{1 \times (k-1)}$ ועברנו למצב $(0, \dots, 1, \dots, 0)_{1 \times (k-1)}$ כאשר 1 הוא במקום ה- j , כלומר $X_j = 1$, אז השינוי ב- Y הוא $\hat{\beta}_j$. במצב הראשון קטגוריית הרפרנס היא 1 (כי כל השאר 0) ואילו במצב השני קטגוריית הרפרנס היא 0 ואילו $X_j = 1$.

9 פירוק לסכום ריבועים

הגדרה: נכליל הגדרות לסכומי ריבועים שונים שהגדרנו עבור רגרסיה פשוטה.

תהי $\mathbb{X}$ מטריצה מגודל $p \times n$ ויהי Y וקטור באורך p , ונניח מודל $Y = \mathbb{X}\beta + \epsilon$, עם

$$\mathbb{1}\{A_1\} = \begin{cases} 0 & \text{not } A_1 \\ 1 & A_1 \end{cases} \quad \text{כלומר: } ^4\text{סימנו כאן ב-} \mathbb{1} \text{ את הפונקציה המציינת.}$$

אומד ריבועים פחותים $\hat{Y} = \mathbb{X}\hat{\beta}$ נגדיר:

$$SSE := \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2 = \|e\|^2$$

$$SST := \sum_{i=1}^n (Y_i - \bar{Y})^2 = \|Y - \bar{Y}\|^2$$

$$SSR := \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \|\hat{Y} - \bar{Y}\|^2$$

סימון: לצורך הנוחות, נסמן $1_{(n)} = (1, \dots, 1)_{1 \times n}$ ונסמן $\bar{Y}_{(n)} = (\bar{Y}, \dots, \bar{Y})_{1 \times n}$.

טענה: כמו במקרה החד-ממדי, מתקיים $SST = SSE + SSR$.

הוכחה: לצורך ההוכחה נראה תחילה שני שוויונים שימשו אותנו:

$$\langle \bar{Y}_{(n)}, e \rangle = \bar{Y}_{(n)}^T e = \sum_{i=1}^n \bar{Y} e_i = \bar{Y} \sum_{i=1}^n e_i = 0$$

$$\langle \hat{Y}, e \rangle = \hat{Y}^T e = \sum_{i=1}^n \hat{Y}_i e_i = 0$$

כדי להראות את השוויונים הללו, נשים לב שבשני המקרים אחד מהווקטורים הוא במרחב $Span(\mathbb{X})$ (כלומר $\bar{Y}_{(n)}, \hat{Y}$ בהתאמה) והאחר הוא במרחב הניצב $Span(\mathbb{X})^\perp$ (כלומר e בשני המקרים). בזאת נסיים, שכן כל מכפלה פנימית של וקטור עם וקטור שניצב לו, היא אפס.

משוויונים אלה נקבל את השוויון הבא:

$$\langle \bar{Y}_{(n)} - \hat{Y}, e \rangle = \langle \bar{Y}_{(n)}, e \rangle - \langle \hat{Y}, e \rangle = 0$$

כלומר $\bar{Y}_{(n)} - \hat{Y}$ ניצב ל- e .

נזכור שלכל זוג וקטורים $u, v \in \mathbb{R}^d$ ניצבים, כלומר $\langle u, v \rangle = 0$, לפי משפט פיתגורס מתקיים $\|u + v\|^2 = \|u\|^2 + \|v\|^2$. אם כך במקרה שלנו נקבל:

$$\begin{aligned} SST &= \|Y - \bar{Y}_{(n)}\|^2 = \|(Y - \hat{Y}) + (\hat{Y} - \bar{Y}_{(n)})\|^2 = \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}_{(n)}\|^2 = SSE + SSR \end{aligned}$$

■

⁵ברור כי $\hat{Y} \in Span(\mathbb{X})$. גם $\bar{Y} \in Span(\mathbb{X})$ כי ראינו שמתקיים $\bar{Y} = \bar{\bar{Y}}$, וברור כי $\bar{\bar{Y}} \in Span(\mathbb{X})$, שכן הוא צירוף לינארי של הווקטורים $\hat{Y}_1, \dots, \hat{Y}_n \in Span(\mathbb{X})$.

מסקנה: כעת ניתן להגדיר מדד לטיב ההתאמה של הרגרסיה גם במקרה הרב-ממדי באופן דומה למקרה החד-ממדי, על ידי:

$$R^2 := 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

הערה: במקרה החד-ממדי מצאנו כי $R^2 = r_{X,Y}^2$ כאשר $r_{X,Y}$ מקדם המתאם של פירסון. במקרה הרב-ממדי הביטוי " $r_{X,Y}$ " אינו מוגדר (כי X מטריצה ולא וקטור).

עם זאת, נזכור שבמקרה החד-ממדי הוכחנו שוויון זה עם הזהות $r_{X,Y}^2 = r_{Y,\hat{Y}}^2$. נשים לב שהביטוי $r_{Y,\hat{Y}}$ מוגדר גם במקרה הרב-ממדי, וניתן להכליל זהות זו גם למקרה הרב-ממדי על ידי $R^2 = r_{Y,\hat{Y}}^2$.

10 הסקה סטטיסטית - השוואת מודלים מקוננים

הגדרה: נדון במודל רגרסיה רב-ממדי נורמלי מהצורה $Y = X\beta + \epsilon$. נתבונן בתת-קבוצה כלשהי $\beta_{j_1}, \dots, \beta_{j_l}$ של רכיבים בווקטור β , ונבדוק את ההשערות:

$$\begin{cases} H_0 : \beta_{j_1} = \dots = \beta_{j_l} = 0 \\ H_1 : \exists_{i \in \{1, \dots, l\}} \beta_{j_i} \neq 0 \end{cases}$$

כלומר, מעוניינים לבדוק את ההשערה שקבוצה מסוימת של רכיבים מהווקטור β היא בעלת השפעה מובהקת על המשתנה Y .

מוטיבציה: מדוע השערה כזאת יכולה להיות מעניינת?

ראשית, אם נבחר כתת-קבוצה את כל רכיבי הווקטור β , אז ההשערה H_0 היא ההשערה שלמודל שלנו $Y = X\beta + \epsilon$ אין כלל יכולת חיזוי. כלומר הווקטור β אינו מייצג כל השפעה של X על Y .

בנוסף, לעתים ישנן קבוצות של משתנים מסבירים שיש ביניהן קשר הגיוני. כך למשל אם מעוניינים לבדוק השפעה של משתנים על רייטינג של תכנית טלוויזיה באמצעות המשתנים של פרסום בעיתון, פרסום באינטרנט ופרסום בטלוויזיה הם בעלי בסיס משותף, וייתכן שנרצה לבדוק את ההשערה שפרסום משלושת הסוגים מסביר את הרייטינג.

היבט נוסף הוא כוח סטטיסטי, כלומר עצמת המבחן. במילים אחרות, נרצה שאם $\beta_j \neq 0$ לאיזה j , אז באמת נגלה זאת על ידי המבחן. כך למשל אם $X_1, \dots, X_{30}$ מייצגים את מספר שעות הלימוד לקראת מבחן, כאשר X_j מייצג j שעות לימוד. כל β_j הוא בעל השפעה מינורית, אולם ברור שלכולם יחד יש השפעה מובהקת על הציון.

סימון: נניח שסידרנו את המקדמים שמצויים תמיד במודל בתחילת הווקטור β , ואת המקדמים שמעוניינים לבדוק את השפעת בשאר הווקטור β . נכתוב $\beta = (\beta^{(1)}, \beta^{(2)})$, כאשר $\beta^{(1)}$ הוא וקטור באורך r המייצג את המקדמים שמצויים תמיד במודל ואילו $\beta^{(2)}$ הוא וקטור באורך $p - r$ המייצג את המקדמים שמעוניינים לבדוק את השפעתם. בהתאמה, נכתוב $X = (X^{(1)}, X^{(2)})$, כאשר $\beta^{(i)}$ הם המקדמים המתאימים למטריצה החלקית $X^{(i)}$.

כעת נוכל לכתוב מחדש את ההשערות שמעוניינים לבדוק:

$$\begin{cases} H_0 : \beta^{(2)} = 0_{p-r} \\ H_1 : \beta^{(2)} \neq 0_{p-r} \end{cases}$$

כאשר $0_{p-r} := (0, \dots, 0)_{1 \times p-r}$.

הגדרה: נתבונן בשני מודלים אפשריים:

$$Y = \mathbb{X}\beta + \epsilon \quad (1)$$

$$Y = \mathbb{X}^{(1)}\beta^{(1)} + \epsilon \quad (2)$$

המודל הראשון מייצג את זה שהמקדמים $\beta^{(2)}$ מייצגים השפעה מובהקת על Y , ואילו המודל השני מייצג את זה שהמקדמים $\beta^{(2)}$ אינם מייצגים השפעה כזאת.

נשים לב כי מודל (2) הוא מקרה פרטי של מודל (1), כי מודל (2) מכיל את המקרה של $\beta = (\beta^{(1)}, \beta^{(2)}) = (0, \beta^{(2)})$. לכן אם סתם נשווה בין שני המודלים נקבל שתמיד המודל הראשון עדיף. זו הסיבה שמודלים אלה מכונים **מודלים מקוננים**.

הגדרה: כדי להתמודד עם הבעיה של הקינון בין שני המודלים, נגדיר את **מבחן יחס הנראות** כדי להשוות בין המודלים. כלומר, נגדיר את הסטטיסטי הבא:

$$LR := \frac{\text{Likelihood}(\mathbb{X}, Y; \hat{\beta}_{\text{mle}}, \hat{\sigma}_{\text{mle}})}{\text{Likelihood}(\mathbb{X}, Y; \hat{\beta}_{\text{mle}}^{(1)}, \hat{\sigma}_{\text{mle}}^{(1)})}$$

כאשר **Likelihood** היא פונקציית הנראות של Y כמשתנה מקרי, במונה תחת הנחת מודל (1) ובמכנה תחת הנחת מודל (2).

הגדרה: יהיו $U_1 \sim \chi_k^2$, $U_2 \sim \chi_m^2$ משתנים מקריים בלתי תלויים, אזי:

$$\frac{U_1/k}{U_2/m} \sim \mathcal{F}_{(k,m)}$$

משפט: במודל רגרסיה $Y = \mathbb{X}\beta + \epsilon$ נורמלי, נסמן $\mathbb{X} = (\mathbb{X}^{(1)}, \mathbb{X}^{(2)})$ כאשר $\mathbb{X}^{(1)}$ מטריצה מגודל $n \times r$ וכן $\mathbb{X}^{(2)}$ מטריצה מגודל $n \times (p-r)$, ובהתאמה נסמן גם $\beta = (\beta^{(1)}, \beta^{(2)})$.

$$\text{עבור בדיקת ההשערות} \begin{cases} H_0 : \beta^{(2)} = 0 \\ H_1 : \beta^{(2)} \neq 0 \end{cases} \text{ מתקיימים שני הדברים הבאים:}$$

1. מבחן יחס הנראות המוכלל שקול למבחן הבא:

$$F := \frac{SSE_1 - SSE/p-r}{SSE/n-p} > C$$

2. תחת H_0 מתקיים $F \sim \mathcal{F}_{(p-r, n-p)}$.

הוכחה:

1. מבחן יחס הנראות המוכלל הוא:

$$\mathbf{LR} := \frac{\text{Likelihood}(\mathbb{X}, y; \hat{\beta}_{\text{mle}}, \hat{\sigma}_{\text{mle}})}{\text{Likelihood}(\mathbb{X}, y; \hat{\beta}_{\text{mle}}^{(1)}, \hat{\sigma}_{\text{mle}}^{(1)})} > \tilde{C}$$

נשים לב כי תחת המודל הנורמלי מתקיים $\hat{\beta}_{\text{mle}} = \hat{\beta}$ (כאשר $\hat{\beta}$ הוא אומדן הריבועים הפחותים), וכן עבור אומדי נראות מירבית במודל נורמלי מתקיים $\hat{\sigma}_{\text{mle}}^2 = SSE/n = \|e\|^2/n$. לכן ניתן לכתוב את המבחן:

$$\mathbf{LR} = \frac{\text{Likelihood}(\mathbb{X}, y; \hat{\beta}, \|e\|^2/n)}{\text{Likelihood}(\mathbb{X}, y; \hat{\beta}^{(1)}, \|e^{(1)}\|^2/n)} > \tilde{C}$$

נחשב ביטוי זה במפורש:

$$\mathbf{LR} = \frac{\frac{1}{(2\pi)^{n/2} \left(\frac{\|e\|^2}{n}\right)^{n/2}} \exp\left(-\frac{1}{2} (Y - \mathbb{X}\hat{\beta})^T \left(\frac{\|e\|^2}{n} I\right)^{-1} (Y - \mathbb{X}\hat{\beta})\right)}{\frac{1}{(2\pi)^{n/2} \left(\frac{\|e^{(1)}\|^2}{n}\right)^{n/2}} \exp\left(-\frac{1}{2} (Y - \mathbb{X}^{(1)}\hat{\beta}^{(1)})^T \left(\frac{\|e^{(1)}\|^2}{n} I\right)^{-1} (Y - \mathbb{X}^{(1)}\hat{\beta}^{(1)})\right)}$$

נצמצם איברים ונכתוב מחדש את מכפלת המטריצות שבאקספוננט, ונקבל:

$$\begin{aligned}
 \mathbf{LR} &= \left(\frac{\|e^{(1)}\|^2}{\|e\|^2} \right)^{n/2} \frac{\exp \left(-\frac{1}{2} \|Y - \mathbb{X}\hat{\beta}\|^2 \frac{n}{\|e\|^2} \right)}{\exp \left(-\frac{1}{2} \|Y - \mathbb{X}^{(1)}\hat{\beta}^{(1)}\|^2 \frac{n}{\|e^{(1)}\|^2} \right)} = \\
 &= \left(\frac{\|e^{(1)}\|}{\|e\|} \right)^n \frac{\exp \left(-\frac{1}{2} \|Y - \hat{Y}\|^2 \frac{n}{\|e\|^2} \right)}{\exp \left(-\frac{1}{2} \|Y - \hat{Y}^{(1)}\|^2 \frac{n}{\|e^{(1)}\|^2} \right)} = \\
 &= \left(\frac{\|e^{(1)}\|}{\|e\|} \right)^n \frac{\exp \left(-\frac{n}{2} \|e\|^2 \frac{n}{\|e\|^2} \right)}{\exp \left(-\frac{1}{2} \|e^{(1)}\|^2 \frac{n}{\|e^{(1)}\|^2} \right)} = \\
 &= \left(\frac{\|e^{(1)}\|}{\|e\|} \right)^n \frac{\exp \left(-\frac{n}{2} \right)}{\exp \left(-\frac{n}{2} \right)} = \left(\frac{\|e^{(1)}\|}{\|e\|} \right)^n
 \end{aligned}$$

כלומר נקבל בסך הכל כי המבחן הוא:

$$\mathbf{LR} = \left(\frac{\|e^{(1)}\|}{\|e\|} \right)^n > \tilde{C}$$

כעת נשים לב כי מתקיים:

$$\frac{SSE_1 - SSE}{SSE} = \frac{SSE_1}{SSE} - 1 = \frac{\|e^{(1)}\|^2}{\|e\|^2} - 1$$

וכעת ניתן להעלות בחזקה, להכפיל ב- $p - n$ ולחלק ב- $p - r$ ולקבל כי אלו מבחנים שקולים. ■

2. כדי להראות כי $F := \frac{SSE_1 - SSE/p - r}{SSE/n - p} \sim \mathcal{F}_{(p-r, n-p)}$ יש להראות כי זו מנה של

שני מ"מ ב"ת מפולגים $\mathcal{X}_{(p-r)}^2, \mathcal{X}_{(n-p)}^2$.

ראינו כבר כי $\frac{SSE}{\sigma^2} \sim \mathcal{X}_{(n-p)}^2$. לכן אם נראה כי $\frac{SSE_1 - SSE}{\sigma^2} \sim \mathcal{X}_{(p-r)}^2$, וכן כי $\frac{SSE_1 - SSE}{\sigma^2}, \frac{SSE}{\sigma^2}$ ב"ת, נסיים את מה שנדרש.

• נראה כי $\frac{SSE_1 - SSE}{\sigma^2}, \frac{SSE}{\sigma^2}$ ב"ת.

נסמן $w := \hat{Y} - \hat{Y}^{(1)}$, כאשר $\hat{Y} = \mathbb{X}\hat{\beta} = P_{\mathbb{X}}Y$ ואילו $\hat{Y}^{(1)} = \mathbb{X}^{(1)}\hat{\beta}^{(2)} = P_{\mathbb{X}^{(1)}}Y$ (עבור $P_{\mathbb{X}^{(1)}}Y$ מטריצת הטלה למרחב הנפרש על ידי עמודות $\mathbb{X}^{(1)}$). תחילה נראה כי w, e הם ב"ת. נחשב:

$$\begin{aligned} w &= \hat{Y} - \hat{Y}^{(1)} = P_{\mathbb{X}}Y - P_{\mathbb{X}^{(1)}}Y = P_{\mathbb{X}}Y - P_{\mathbb{X}^{(1)}}P_{\mathbb{X}}Y = \\ &= (I - P_{\mathbb{X}^{(1)}})P_{\mathbb{X}}Y = (I - P_{\mathbb{X}^{(1)}})\hat{Y} = (I - P_{\mathbb{X}^{(1)}})\mathbb{X}\hat{\beta} \end{aligned}$$

כאשר השוויון השלישי נובע מכך שמתקיים $P_{\mathbb{X}^{(1)}}P_{\mathbb{X}} = P_{\mathbb{X}^{(1)}}$, ושאר השוויונים זהותיים. כעת נשים לב כי $(I - P_{\mathbb{X}^{(1)}})\mathbb{X}$ היא מטריצה קבועה כלשהי. כמו כן נזכור כי $\hat{\beta}, e$ מ"מ ב"ת, ולכן גם w, e מ"מ ב"ת, שכן w הוא $\hat{\beta}$ מוכפל בקבוע. כעת נשים לב כי $SSE = \|e\|^2$, וכן נשים לב שמתקיים:

$$\|e^{(1)}\|^2 = \|Y - \hat{Y}^{(1)}\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \hat{Y}^{(1)}\|^2 = \|e\|^2 + \|\hat{Y} - \hat{Y}^{(1)}\|^2 = \|e\|^2 + \|w\|^2$$

כאשר השוויון השני הוא ממשפט פיתגורס. מכאן נובע:

$$SSE_1 - SSE = \|e^{(1)}\|^2 - \|e\|^2 = \|w\|^2$$

מכאן כי $\frac{SSE_1 - SSE}{\sigma^2}$ הוא פונקציה של w וכן $\frac{SSE}{\sigma^2}$ הוא פונקציה של e , ולכן מהיות w, e ב"ת נובע כי גם $\frac{SSE_1 - SSE}{\sigma^2}, \frac{SSE}{\sigma^2}$ ב"ת.⁶ נראה כי $\frac{SSE_1 - SSE}{\sigma^2} \sim \chi^2_{(p-r)}$. נשים לב שמתקיים:

$$\begin{aligned} w &= \hat{Y} - \hat{Y}^{(1)} = (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})Y = (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})(\mathbb{X}\beta + \epsilon) = \\ &= (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\left((\mathbb{X}^{(1)}, \mathbb{X}^{(2)})\left(\beta^{(1)}, \beta^{(2)}\right)^T + \epsilon\right) = \\ &= (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})(\mathbb{X}^{(1)}\beta^{(1)} + \mathbb{X}^{(2)}\beta^{(2)} + \epsilon) = \\ &= (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\mathbb{X}^{(1)}\beta^{(1)} + (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\mathbb{X}^{(2)}\beta^{(2)} + (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\epsilon = \\ &= (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\epsilon \end{aligned}$$

כאשר השוויון בין השורה הרביעית והחמישית נובע כי:

$$(P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\mathbb{X}^{(1)} = 0$$

$$(P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\mathbb{X}^{(2)} = 0$$

כלומר קיבלנו את השוויון:

$$w = (P_{\mathbb{X}} - P_{\mathbb{X}^{(1)}})\epsilon$$

⁶ באופן כללי, אם U, V זוג מ"מ ב"ת, אז גם $f(U), g(V)$ מ"מ ב"ת, עבור כל פונקציות f, g .

מכאן נובע כי:

$$\begin{aligned} SSE_1 - SSE &= \|w\|^2 = \|(P_X - P_{X(1)})\epsilon\|^2 = \\ &= [(P_X - P_{X(1)})\epsilon]^T (P_X - P_{X(1)})\epsilon = \epsilon^T (P_X - P_{X(1)})^T (P_X - P_{X(1)})\epsilon = \\ &= \epsilon^T (P_X - P_{X(1)})\epsilon \end{aligned}$$

כאשר השוויון בשורה האחרונה נובע כי $P_X - P_{X(1)}$ היא מטריצת הטלה על המרחב הניצב $\text{Span}(X^{(1)})^\perp$, ובאופן כללי לכל מטריצת הטלה P_A מתקיים $P_A P_A = P_A$. מכאן נמשיך באותו אופן בו הוכחנו בעבר כי $\frac{n-p}{\sigma^2} s^2 \sim \chi^2_{n-p}$. כלומר, נשתמש במשפט הפירוק הספקטרלי. בהתאם למשפט זה, ניתן לפרק את המטריצה הסימטרית $P_X - P_{X(1)}$ לצורה $P_X - P_{X(1)} = U\Lambda U^T$, כאשר U היא מטריצה אורתוגונלית (כלומר $U^T U = I$) שהעמודות שלה הם הווקטורים העצמיים של $P_X - P_{X(1)}$, וכן Λ היא מטריצה שכולה אפסים חוץ מהאלכסון הראשי שלה, שמכיל את הערכים העצמיים של $P_X - P_{X(1)}$ (כלומר היא מהצורה $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$). באותו אופן מראים שיש בדיוק $p - r$ פעמים 1 באלכסון הראשי הנ"ל, ומקבלים את השוויון:

$$\|w\|^2 = \sum_{i=1}^{p-r} \epsilon_i^2$$

שממנו נובע השוויון:

$$\frac{\|w\|^2}{\sigma^2} = \sum_{i=1}^{p-r} \frac{\epsilon_i^2}{\sigma^2}$$

אבל $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ ולכן $\frac{\epsilon_i}{\sigma} \sim \mathcal{N}(0, 1)$, ומכאן כי $\frac{\|w\|^2}{\sigma^2}$ הוא סכום של $p - r$ מ"מ נורמלים סטנדרטיים, ולכן $\frac{SSE_1 - SSE}{\sigma^2} = \frac{\|w\|^2}{\sigma^2} \sim \chi^2_{(p-r)}$. ■

10.1 מבחן F מול מבחן t

נשים לב כי עבור בדיקת ההשערות $\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$ ניתן להשתמש במבחן F על ידי

$$\tilde{C} < \frac{SSE_1 - SSE/1}{SSE/n-p} \sim F_{(1, n-p)} \quad \text{וניתן להשתמש במבחן } t \text{ על ידי } C < \frac{\hat{\beta}_j}{s\sqrt{(X^T X)^{-1}_{jj}}} \sim t_{(n-p)}$$

לא נוכיח זאת, אולם עבור המקרה המיוחד של בדיקת השערות מסוג זה, שני המבחנים הללו שקולים. כלומר, מבחן F מכליל את מבחן t .

הערה: באופן כללי מבחן F בודק את המובהקות של קבוצת משתנים, אולם הוא לא אומר שום דבר על כל אחד מהמשתנים שבקבוצה.

עקרונית ייתכן מצב בו מבחן F יצביע על השפעה מובהקת של קבוצת משתנים, אבל לא יצביע על השפעה מובהקת בהפעלה שלו על כל אחד מהמשתנים בקבוצה. מצב כזה עלול להיווצר במקרה בו אין מספיק נתונים במדגם.

Adjust R^2 11

הגדרנו בעבר מדד לטיב ההתאמה $R^2 := 1 - \frac{SSE}{SST}$. אולם נשים לב שמדד זה אינו מאפשר השוואה בין מודלים בעלי ממדים שונים. כך למשל בהשוואה בין מודל $Y = \mathbb{X}\beta + \epsilon$ לבין מודל $Y = \mathbb{X}^{(1)}\beta^{(1)} + \epsilon$, תמיד יתקיים $SSE_1 \geq SSE$ ולכן $R^{2(1)} = 1 - \frac{SSE_1}{SST} \leq 1 - \frac{SSE}{SST} = R^2$. לפיכך נרצה להגדיר מדד לטיב התאמה שיאפשר להשוות את טיב ההתאמה בין מודלים בעלי ממדי שונה.

הגדרה:

$$\bar{R}^2 := 1 - \frac{SSE/n-p}{SST/n-1}$$

טענה: מתקיים זהותית:

$$\bar{R}^2 = R^2 - (1 - R^2) \frac{p-1}{n-p}$$

וכן:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-p}$$

טענה: מתקיים תמיד $R^2 \geq \bar{R}^2$.

הוכחה: נובע מידית מהזהויות הקושרות בין $\bar{R}^2$ לבין R^2 .

12 בדיקת הנחות המודל

נתבונן במודל רגרסיה נורמלי $Y = \mathbb{X}\beta + \epsilon$ כאשר $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ שרכיביו ב"ת. לעתים נחשוך בהנחה כי $\epsilon \sim \mathcal{N}$, ולשם כך נרצה לאמוד את וקטור השגיאות $\epsilon = Y - X^T \hat{\beta}$. נתייחס אל $e = Y - \hat{Y}$ כאל אומד של ϵ .

טענה: מתקיים $e \sim \mathcal{N}(0, \sigma^2 (I - P_{\mathbb{X}}))$.

הוכחה: נתייחס אל $e_i = Y_i - \hat{Y}_i$ כאל אומד לשגיאה ϵ_i , ונרצה לחשב את המומנטים של e_i . לשם כך נשים לב כי:

$$e = Y - \hat{Y} = \mathbb{X}\beta + \epsilon - \mathbb{X}\hat{\beta} = \mathbb{X}(\beta - \hat{\beta}) + \epsilon$$

מכאן נובע כי:

$$E[e] = E[\mathbb{X}(\beta - \hat{\beta}) + \epsilon] = \mathbb{X}(E[\hat{\beta}] - \beta) + E[\epsilon] = 0 + 0 = 0$$

$$\begin{aligned}
\text{Var}(e) &= \text{Var}(Y - \hat{Y}) = \text{Var}(\mathbb{X}\beta + \epsilon - \mathbb{X}\hat{\beta}) = \text{Var}(\epsilon - \mathbb{X}\hat{\beta}) = \\
&= \text{Var}(\epsilon - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY) = \text{Var}(\epsilon - P_{\mathbb{X}}Y) = \text{Var}(\epsilon - P_{\mathbb{X}}(\mathbb{X}\beta + \epsilon)) = \\
&= \text{Var}(\epsilon - P_{\mathbb{X}}\mathbb{X}\beta - P_{\mathbb{X}}\epsilon) = \text{Var}(\epsilon - P_{\mathbb{X}}\epsilon) = \text{Var}((I - P_{\mathbb{X}})\epsilon) = \\
&= (I - P_{\mathbb{X}}) \underbrace{\text{Var}(\epsilon)}_{=\sigma^2} (I - P_{\mathbb{X}})^T = \sigma^2 (I - P_{\mathbb{X}})(I - P_{\mathbb{X}})^T = \sigma^2 (I - P_{\mathbb{X}})
\end{aligned}$$

כאשר השוויון האחרון נובע מתכונות של מטריצת הטלה P_A .⁷

הגדרה: נשים לב כי התפלגות e תלויה בשונות $I - P_{\mathbb{X}}$ התלויה בעצמה בתצפיות $\mathbb{X}$. לשם כך נתקן את השאריות. תחילה נשים לב כי הווקטור $\tilde{e}$ המוגדר על ידי $\tilde{e}_i := \frac{e_i}{\sigma\sqrt{(I - P_{\mathbb{X}})_{ii}}}$ מקיים $\tilde{e} \sim \mathcal{N}(0, 1)$. אך נשים לב כי σ אינה ידועה, לכן נציב במקום זאת את s . כלומר, נגדיר את וקטור השאריות המתוקנות e^* להיות:

$$e_i^* := \frac{e_i}{s\sqrt{(I - P_{\mathbb{X}})_{ii}}}$$

נקבל כי $e^* \sim t_{(n-p)}$, ונזכור כי זה קירוב יחסית טוב להתפלגות הנורמלית הסטנדרטית כאשר n גדול בהרבה מ- p . כמו כן מתקיים כי השונות של e^* אינה תלויה בתצפיות.

שיטת $QQ - \text{plot}$: נניח כי נתון וקטור השאריות המתוקנות e^* . נסמן באופן כללי את האחוזון α -של מ"מ Z על ידי q_α . כלומר $P(Z \leq q_\alpha) = \alpha$ (עבור $0 \leq \alpha \leq 1$).⁸ נסמן את "סטטיסטי הסדר" על ידי $e_{(1)}^* \leq e_{(2)}^* \leq \dots \leq e_{(n)}^*$. כעת נצייר מערכת צירים, כאשר ציר ה- x הוא האחוזונים של ההתפלגות הנורמלית (z_α) ועל ציר ה- y הוא האחוזונים האמפיריים של המדגם (q_α). ככל שהנקודות שנצייר יהיו קרובות יותר לישר $y = x$, תתחזק ההנחה שאכן מדובר בהתפלגות נורמלית.

13 ייצוב השונות

הגדרה: נרצה להניח כי השונויות שוות (כלומר $\text{Var}(\epsilon_i) = \text{Var}(\epsilon_j)$), בעוד שזה לא המצב תמיד. לשם כך נחפש טרנספורמציה מייצבת שונות, כלומר טרנספורמציה g , כך שעבור $\tilde{Y} := g(Y)$ הנחת השונויות השוות תתקיים.

הערה: נשים לב כי $\text{Var}(Y_i) = \text{Var}(X_i\beta + \epsilon_i) = \text{Var}(\epsilon_i)$.

הערה: אחת מהנחות המודל היא כי $\text{Var}(Y_i)$ הוא בלתי תלוי בתוחלת $E[Y_i] := \mu_i$. במקרה שהנחה זו אינה מתקיימת, כלומר הם כן תלויים, נרצה למצוא תלות פונקציונלית כלשהי $\text{Var}(Y_i) = f(\mu_i)$.

סימון: נסמן $E[Y_i | X_i] := \mu_i$ וכן $\text{Var}(Y_i | X_i) := f(\mu_i)$ עבור פונקציה f כלשהי.

⁷שכן $I - P_A$ היא עצמה מטריצת הטלה למרחב הניצב, וכן כי באופן כללי $P_A^2 = P_A$.
⁸בהתפלגות הנורמלית הסטנדרטית נהוג לסמן $q_\alpha = z_\alpha$.

טענה: יהי Z מ"מ כלשהו. נסמן $E[Z] = \mu$ וכן $Var(Z) = \sigma^2$. תהי גם g טרנספורמציה כלשהו. אזי עבור המ"מ $g(Z)$ מתקיים הקירוב:

$$E[g(Z)] \approx g(\mu) + \frac{1}{2}g''(\mu)\sigma^2$$

$$Var(g(Z)) \approx g'(\mu)^2\sigma^2$$

הוכחה: נשתמש בפיתוח טיילור סביב μ , ונקבל:

$$g(z) = g(\mu) + g'(\mu)(z - \mu) + \frac{1}{2}g''(\mu)(z - \mu)^2 + o((z - \mu)^3)$$

כעת נחשב את התוחלת בנקודה Z :

$$\begin{aligned} E[g(Z)] &= \\ &= E[g(\mu)] + E[g'(\mu)(Z - \mu)] + E\left[\frac{1}{2}g''(\mu)(Z - \mu)^2\right] + E\left[o((Z - \mu)^3)\right] = \\ &= g(\mu) + g'(\mu)\underbrace{(E[Z] - \mu)}_{=0} + \frac{1}{2}g''(\mu)E[(Z - \mu)^2] + E\left[o((Z - \mu)^3)\right] = \\ &= g(\mu) + \frac{1}{2}g''(\mu)\sigma^2 + o((Z - \mu)^3) \end{aligned}$$

ונחשב את השונות בנקודה Z :

$$\begin{aligned} Var(g(Z)) &= E[g(Z)^2] - E[g(Z)]^2 = \\ &= E\left[\left(g(\mu) + g'(\mu)(Z - \mu) + \frac{1}{2}g''(\mu)(Z - \mu)^2 + o((Z - \mu)^3)\right)^2\right] - \\ &\quad - \left(g(\mu) + \frac{1}{2}g''(\mu)\sigma^2\right)^2 + o((Z - \mu)^3) = \\ &= E\left[g(\mu)^2 + g'(\mu)^2(Z - \mu)^2 + 2g(\mu)g'(\mu)(Z - \mu) + 2g(\mu)\frac{1}{2}g''(\mu)(Z - \mu)^2\right] - \\ &\quad - \left[g(\mu)^2 + \frac{1}{4}g''(\mu)^2\sigma^4 + g(\mu)g''(\mu)\sigma^2\right] + o((Z - \mu)^3) = \\ &= g'(\mu)^2\sigma^2 - \frac{1}{4}g''(\mu)^2\sigma^4 + o((Z - \mu)^3) = \\ &= g'(\mu)^2\sigma^2 - \frac{1}{4}g''(\mu)^2E[(Z - \mu)^2]^2 + o((Z - \mu)^3) = \\ &= g'(\mu)^2\sigma^2 + o((Z - \mu)^3) \end{aligned}$$

■

מסקנה: יהי $Y | \mathbb{X}$ המקיים $E[Y_i | X_i] = \mu_i$, ונניח שקיימת פונקציה f המקיימת $Var(Y_i | X_i) = f(\mu_i)$. אזי קיימת טרנספורמציה מייצבת שונות g , המתקבלת על ידי:

$$g(t) := \tilde{c} \cdot \int \frac{1}{\sqrt{f(t)}} dt$$

עבור $\tilde{c} \in \mathbb{R}$ קבוע כלשהו.

הוכחה: אנו מחפשים טרנספורמציה g שעבורה:

$$Var(g(Y_i) | X_i) = f(\mu_i) = c$$

לכל $i = 1, \dots, n$ עבור $c \in \mathbb{R}$ קבוע.

מהטענה שהראינו לעיל נובע כי מתקיים באופן כללי:

$$Var(g(Y_i) | X_i) = g'(\mu_i)^2 Var(Y_i | X_i) = g'(\mu_i)^2 f(\mu_i)$$

ולכן הדרישה היא:

$$g'(\mu_i)^2 f(\mu_i) = c$$

מכאן נובע כי:

$$g'(\mu_i) = \sqrt{\frac{c}{f(\mu_i)}}$$

לכן נוכל לבחור:

$$g(t) := \sqrt{c} \cdot \int \frac{1}{\sqrt{f(t)}} dt$$

ונקבל טרנספורמציה המקיימת את הנדרש עבור c (עד כדי הוספת קבוע). ■

דוגמה: (רגרסיה פואסונית) נניח כי $X_1, \dots, X_p$ הן תכונות של כלי רכב, וכי Y הוא מספר התאונות בהן הם מעורבים, ונניח כי $Y | \mathbb{X} \sim Poiss\left(\sum_{j=1}^p \beta_j X_j\right)$.

מתקיים כי $E[Y_i | X_i] = Var(Y_i | X_i) = \mu_i$ (מתכונות ההתפלגות הפואסונית), ולכן הקשר בין התוחלת והשונות הוא $Var(Y_i | X_i) = f(\mu_i) = \mu_i$. מכאן כי טרנספורמציה מייצבת שונות g היא:

$$g(t) = \sqrt{c} \int \sqrt{\frac{1}{f(t)}} dt = \sqrt{c} \int \sqrt{\frac{1}{t}} dt = \sqrt{c} \int t^{-1/2} dt = 2\sqrt{ct}$$

כלומר נפעיל את הטרנספורמציה $g(t) = \sqrt{t}$ (עד כדי קבוע) על המדגם, ונקבל כי עבור $g(Y_i) = \sqrt{Y_i}$ מתקיימת הנחת השונויות השוות.

14 מדדים לחריגות של תצפית

מבוא: במסגרת הנחת המודל שלנו לגבי המדגם, לעתים מתוך המדגם ניתן להבחין בתצפיות מסוימות שהן חריגות מספיק כדי שניח כי הן בעלות מאפיינים שונים. כלומר שהנחות המודל אינן נכונות לגביהן, ולכן יש לשקול להסיר אותן מהמדגם. כדי להגדיר חריגות כאלה, נגדיר שני מדדים אפשריים. המדד הראשון ("הנפה") בודק עד כמה התצפית ה- i חורגת מהמרכז ועד כמה הכיוון שלה שונה, והמדד השני ("מרחק קוק") בודק עד כמה התצפית ה- i משפיעה על $\hat{\beta}$.

14.1 הנפה

הגדרה: עבור מדגם $(\mathbb{X}, Y)$, **ההנפה** (Leverage) של התצפית ה- i היא $h_{ii} := (P_{\mathbb{X}})_{ii}$ כאשר $P_{\mathbb{X}} := \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$ היא מטריצת ההטלה (שממדיה $n \times n$).

טענה: עבור רגרסיה לינארית פשוטה עם חותך, ההנפה היא $h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}$.

הוכחה: נניח כי $\bar{X} = 0$ והשלמת ההוכחה עבור $\bar{X}$ כלשהו תושאר כתרגיל. נשים לב כי במקרה זה המטריצה $\mathbb{X}$ היא מהצורה $(1_{n \times 1} | X)$ עבור $X \in \mathbb{R}$. כעת נחשב לפי הגדרת הנפה:

$$\begin{aligned} h_{ii} &= (P_{\mathbb{X}})_{ii} = \left(X (X^T X)^{-1} X^T \right)_{ii} = \\ &= \left((1_{n \times 1} | X) \left(\begin{pmatrix} 1_{1 \times n} \\ X^T \end{pmatrix} (1_{n \times 1} | X) \right)^{-1} \begin{pmatrix} 1_{1 \times n} \\ X^T \end{pmatrix} \right)_{ii} = \\ &= \left((1_{n \times 1} | X) \left(\begin{matrix} n & \sum_{j=1}^n X_j \\ \sum_{j=1}^n X_j & \sum_{j=1}^n X_j^2 \end{matrix} \right)^{-1} \begin{pmatrix} 1_{1 \times n} \\ X^T \end{pmatrix} \right)_{ii} = \\ &= \left((1_{n \times 1} | X) \begin{pmatrix} n & 0 \\ 0 & \sum_{j=1}^n X_j^2 \end{pmatrix}^{-1} \begin{pmatrix} 1_{1 \times n} \\ X^T \end{pmatrix} \right)_{ii} = \\ &= \left((1_{n \times 1} | X) \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{j=1}^n X_j^2} \end{pmatrix} \begin{pmatrix} 1_{1 \times n} \\ X^T \end{pmatrix} \right)_{ii} = \frac{1}{n} + \frac{X_i^2}{\sum_{j=1}^n X_j^2} \end{aligned}$$

■ וקיבלנו את הנוסחה המבוקשת תחת ההנחה $\bar{X} = 0$.

טענה: עבור רגרסיה לינארית פשוטה עם חותך, מתקיים $\bar{h} = \frac{2}{n}$.

הוכחה: נחשב:

$$\begin{aligned} \bar{h} &= \frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right] = \\ &= \frac{1}{n} \left[1 + \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right] = \frac{1}{n} [1 + 1] = \frac{2}{n} \end{aligned}$$

■

טענה: עבור רגרסיה לינארית מרובה, מתקיים $\bar{h} = \frac{p}{n}$.

הוכחה: באופן כללי עבור מטריצה $A_{n \times p} = (a_{ij})_{n \times p}$, מגדירים $\text{Trace}(A) = \sum_{i=1}^n a_{ii}$ (סכום איברי האלכסון הראשי). ידוע כי באופן כללי $\text{Trace}(AB) = \text{Trace}(BA)$.
 כעת נחשב:

$$\begin{aligned}\bar{h} &= \frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{1}{n} \text{Trace}(P_{\mathbb{X}}) = \frac{1}{n} \text{Trace}\left(\mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T\right) = \\ &= \frac{1}{n} \text{Trace}\left(\mathbb{X}^T \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1}\right) = \frac{1}{n} \text{Trace}(I_{p \times p}) = \frac{p}{n}\end{aligned}$$

כאשר השוויון השני נובע מכך ש- h_{ii} הוא האיבר ה- i באלכסון הראשי של המטריצה $P_{\mathbb{X}}$, והשוויון השלישי נובע מהתכונה של Trace שהזכרנו. ■

הערה: אין מדד חד-משמעי לחריגות. נהוג להתייחס לתצפית i כחריגה אם $h_{ii} > \frac{2p}{n}$.

הערה: ההנפה אינה מודדת רק את המרחק של תצפית מהממוצע, אלא משקללת את הכיוון של התצפית ביחס לשאר התצפיות.

כלומר, ייתכן שעבור שתי תצפיות (X_i, Y_i) , (X_j, Y_j) יתקיים כי:

$$\|X_i - \bar{X}\| < \|X_j - \bar{X}\|$$

ובמקביל גם:

$$h_{ii} > h_{jj}$$

14.2 מרחק קוק

הגדרה: עבור מדגם $(\mathbb{X}, Y)$, **מרחק קוק** (Cook's distance) של התצפית ה- i הוא:

$$D_i := \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T \mathbb{X}^T \mathbb{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{p \cdot s^2}$$

כאשר $\hat{\beta}_{(-i)}$ הוא אומד הריבועים הפחותים המתקבל עבור המדגם בלי התצפית ה- i .

נוסחה: נשים לב כי:

$$\begin{aligned}(\hat{\beta} - \hat{\beta}_{(-i)})^T \mathbb{X}^T \mathbb{X} (\hat{\beta} - \hat{\beta}_{(-i)}) &= \\ &= \left(\mathbb{X}(\hat{\beta} - \hat{\beta}_{(-i)})\right)^T \left(\mathbb{X}(\hat{\beta} - \hat{\beta}_{(-i)})\right) = \left\|\mathbb{X}(\hat{\beta} - \hat{\beta}_{(-i)})\right\|^2\end{aligned}$$

ולכן:

$$D_i = \frac{\left\|\mathbb{X}(\hat{\beta} - \hat{\beta}_{(-i)})\right\|^2}{ps^2} = \frac{\left\|\mathbb{X}\hat{\beta} - \mathbb{X}\hat{\beta}_{(-i)}\right\|^2}{ps^2} = \frac{\left\|\hat{Y} - \hat{Y}_{(-i)}\right\|^2}{ps^2}$$

הערה: נשים לב כי $Var(\hat{\beta}) = (\mathbb{X}^T \mathbb{X})^{-1} \sigma^2$ ולכן $\hat{Var}(\hat{\beta}) = (\mathbb{X}^T \mathbb{X})^{-1} s^2$

לכן ניתן להתייחס למרחק קוק כאל הנורמה של ההפרש $\|\hat{\beta} - \hat{\beta}_{(-i)}\|$ מנורמלת באומד לשונות, כאשר במכנה הוספנו p כדי לתת משקל מתאים לרכיב ה- i מתוך p הרכיבים.

הגדרה: תהי M מטריצה סימטרית וחיובית לחלוטין כלשהי. **מרחק מהלנוביס** (Mahalanobis) ביחס ל- M , לכל שני וקטורים $\vec{v}_1, \vec{v}_2$ מוגדר להיות:

$$d_M(\vec{v}_1, \vec{v}_2) := (\vec{v}_1 - \vec{v}_2)^T M^{-1} (\vec{v}_1 - \vec{v}_2)$$

הערה: נשים לב כי מרחק קוק הוא מרחק מהלנוביס ביחס למטריצה $M = ps^2 \mathbb{X}^T \mathbb{X}$.

14.2.1 התפלגות מרחק קוק

משפט: לכל $\vec{z} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ מממד p , מתקיים:

$$d_\Sigma(\vec{z}, \vec{\mu}) := (\vec{z} - \vec{\mu})^T \Sigma^{-1} (\vec{z} - \vec{\mu}) \sim \chi_{(p)}^2$$

סקיצה להוכחה: נשים לב שעבור $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ הטענה ברורה מאליה, שכן זה סכום של ריבועי משתנים מקריים נורמליים סטנדרטיים.

עבור Σ כללית, באמצעות פירוק ספקטרלי $\Sigma = V^T \Lambda V$ ניתן שוב להביע את הביטוי הנ"ל כסכום של משתנים מקריים נורמליים סטנדרטיים. ■

מסקנה: ידוע כי $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$ ומכאן כי:

$$(\hat{\beta} - \beta)^T \frac{1}{\sigma^2} \mathbb{X}^T \mathbb{X} (\hat{\beta} - \beta) \sim \chi_{(p)}^2$$

מתקיים $E[\chi_{(p)}^2] = p$, ולכן נקבל:

$$E \left[\frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T \mathbb{X}^T \mathbb{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{p \cdot \sigma^2} \right] = 1$$

אולם σ אינו ידוע, ולכן נציב במקומו את האומד s , ונקבל בקירוב:

$$E[D_i] = E \left[\frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T \mathbb{X}^T \mathbb{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{p \cdot s^2} \right] \approx 1$$

מסקנה: מכאן נובע שבקירוב D_i מתפלג $\mathcal{F}_{(p, n-p)}$, ועל ידי כך להשתמש באחוזון ה- $1 - \alpha$ של התפלגות זו כסף המודד את החריגות של התצפית ה- i .

הערה: צורה נוספת למדוד חריגות של תצפית i כלשהי, היא לבדוק האם $D_i > c/n$ עבור קבוע c כלשהו. לעתים נהוג לקחת $c = 4$.

15 מדדים למולטי-קוליניאריות של משתנה

ישנן כמה דרכים לבדוק האם משתנה מסוים מתוך p המשתנים המסבירים, מתואם לינארית עם המשתנים האחרים.

1. **דרך ראשונה: קורלציה.** ניתן לחשב את $\rho_{ij} := \text{Corr}(X_i, X_j)$ עבור כל $i, j = 1, \dots, p$, ולבדוק האם נמצאת קורלציה גבוהה במיוחד.

החיסרון של דרך זו היא שהיא בודקת קשר בין זוגות משתנים, אך לא קשרים לינאריים מורכבים. ייתכן כי משתנה X_i לא יהיה מתואם עם משתנה X_j ולא עם משתנה X_k , ובמקביל שהוא יהיה מתואם עם צירוף לינארי של X_j, X_k .

2. **דרך שנייה: רגרסיה.** ניתן לבצע רגרסיה של כל משתנה X_i ביחס לשאר המשתנים $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p$. כלומר, ברגרסיה זו אנו מתייחסים אל X_i כמשתנה מוסבר על ידי שאר $p-1$ המשתנים.

כך מבצעים p רגרסיות כאלה, ומתקבלים $R_1^2, \dots, R_p^2$ עבור כל הרגרסיות. ניתן להגדיר $Tol_i := 1 - R_i^2$, וכאשר Tol_i קרוב ל-0 זה אומר שקיים קשר לינארי חזק בין X_i לבין שאר המשתנים.

3. **דרך שלישית: מדד גלובלי.** מדד זה לא יבדוק האם משתנה X_i מסוים הוא בעל קשר לינארי חזק לשאר המשתנים, אלא יבדוק האם במודל שלנו באופן כללי קיים איזשהו קשר כזה.

נשים לב שבמצב בו קיימת תלות לינארית מלאה, כלומר X_i הוא צירוף לינארי של שאר המשתנים, אז המטריצה $\mathbb{X}$ היא מטריצה סינגולרית, כלומר $\text{rank}(\mathbb{X}) < p$. אם כך נפתח מדד שיבדוק עד כמה המטריצה $\mathbb{X}$ קרובה להיות סינגולרית. ככל שהמטריצה תהיה קרובה לסינגולרית, נדע שקיימת מולטי-קוליניאריות במודל שלנו.

נשים לב שלכל מטריצה $A \in \mathcal{M}_{p \times p}(\mathbb{R})$, אם דרגתה לא מלאה, כלומר $\text{rank}(A) < p$, אז קיים וקטור $\vec{u} \in \mathbb{R}^p$ כך שמתקיים $A\vec{u} = 0$, כי אחד מווקטורי העמודה תלוי לינארית באחרים, ולכן וקטור המקדמים של תלות לינארית זו הוא $\vec{u}$ המקיים זאת.

נסמן $S := \{\vec{w} \in \mathbb{R}^p \mid \|\vec{w}\| = 1\}$, ונגדיר את המדד הבא:

$$\text{Cond}(\mathbb{X}) := \frac{\max_{\vec{u} \in S} \|\mathbb{X}\vec{u}\|}{\min_{\vec{v} \in S} \|\mathbb{X}\vec{v}\|}$$

נשים לב שמדד זה מקיים $0 \leq \text{Cond}(\mathbb{X}) \leq \infty$. הוא יכול לקבל גם את הערך ∞ , שכן אם $\text{rank}(\mathbb{X}) < p$ אז אכן קיים $\vec{v} \in S^1$ שעבורו $\|\mathbb{X}\vec{v}\| = 0$.

כלומר, ככל שהערך של $\text{Cond}(\mathbb{X})$ יהיה גבוה יותר, נסיק שהמודל שלנו מכיל מולטי-קוליניאריות רבה יותר.

משפט: נסמן את הערכים העצמיים המקסימלי והמינימלי של המטריצה $\mathbb{X}^T \mathbb{X}$ על ידי $\lambda_{\min}, \lambda_{\max}$. לכל $\vec{u} \in S^1$ מתקיימים הדברים הבאים:

(א)

$$\sqrt{\lambda_{\min}} \leq \text{Cond}(\mathbb{X}) \leq \sqrt{\lambda_{\max}}$$

(ב)

$$\text{Cond}(\mathbb{X}) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

16 אינטראקציות

הגדרה: בהינתן מודל רגרסיה $Y = \mathbb{X}\beta + \epsilon$, קשר פנימי בין חלק מהמשתנים המסבירים $\{X_1, \dots, X_p\}$ נקרא **אינטראקציה**.

נדון באינטראקציות מצורה של מכפלה. כלומר למשל $X_1 \cdot X_2$ היא אינטראקציה זוגית, ולמשל $X_1 \cdot X_2 \cdot X_3$ היא אינטראקציה משולשת.

נתעניין בהשפעה של הוספת אינטראקציות למודל הרגרסיה. כלומר כיצד הוספת משתנה חדש $X_{p+1} = X_1 \cdot X_2$ משנה את מודל הרגרסיה.

הערה: ייתכן כי בין X_1, X_2 אין כלל מולטי-קולינאריות, ובכל זאת הוספת האינטראקציה $X_1 \cdot X_2$ למודל תשפר אותו.

דוגמה: נחשוב על מודל $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, כאשר X_1 הוא משתנה בינארי ואילו X_2 משתנה רציף.

ייתכן כי X_2 לא מסביר בכלל את Y , אולם הכפלה שלו במשתנה הבינארי X_1 תשאיר ממנו רק ערכים המסבירים את Y .

הערה: גם לאחר תוספת אינטראקציות מודל הרגרסיה נשאר בעל אותו מאפיינים. כך למשל אם מעוניינים להוסיף אינטראקציה $X_i \cdot X_j$ כלשהי, אז מגדירים מטריצה של מודל מממד $p + 1$ על ידי:

$$\tilde{\mathbb{X}} = (X_1, \dots, X_p, X_i \cdot X_j)$$

ומבצעים רגרסיה כמו שמבצעים בדרך כלל.

הערה: ניתן היה לחשוב שבמצב אידאלי כדאי לבדוק את כל האינטראקציות האפשריות. כלומר לבחון את המודל:

$$Y = \sum_{j=1}^p \beta_j X_j + \sum_{\substack{k, l = 1 \\ k < l}}^p \beta_{kl} X_k \cdot X_l + \epsilon$$

אולם נשים לב כי לא בהכרח מתקבלת מטריצה $\mathbb{X}$ מדרגה מלאה, שכן מספר כל האינטראקציות הוא $\binom{p}{2} = \frac{p(p-1)}{2}$, ולכן $\text{rank}(\mathbb{X}) = p + \frac{p(p-1)}{2}$. נזכור שתמיד $\text{rank}(\mathbb{X}) \leq n$ (כי דרגת העמודות שווה לדרגת השורות), ולכן ייתכן מצב שנקבל מטריצה בעלת $p + \frac{p(p-1)}{2}$ עמודות, אבל שדרגתה $\text{rank}(\mathbb{X}) \leq n < p + \frac{p(p-1)}{2}$, ולכן $\mathbb{X}^T \mathbb{X}$ אינה הפיכה ולא ניתן לחשב אומד ריבועים פחותים.

17 בחירת מודל

בהינתן משתנים $Y, X_1, \dots, X_p$ עלינו לקבוע מהו המודל המתאים הקושר בין Y לבין $X_1, \dots, X_p$. לשם עלינו להחליט איזו קבוצת מודלים כדאי לנסות (עם כל המשתנים או עם חלקם? עם או בלי אינטראקציות? אם כן, אלו אינטראקציות כדי להכניס? וכדומה), ולאחר מכן להחליט מתוך קבוצת המודלים שבחרנו איזה מודל הוא הטוב ביותר. לשם כך נציג שלושה אלגוריתמים אפשריים:

• אלגוריתם ראשון - בחירה קדימה

- נתחיל עם מודל המכיל חותך בלבד. כלומר $Y = \beta_0 + \epsilon$.
- לכל $j = 1, \dots, p$ נבצע:
 - * נגדיר מודל עם המשתנה x_j . כלומר $Y = \beta_0 + \beta_j x_j + \epsilon$.
 - * נבצע T -test להשערה $H_0 : \beta_j = 0$ ונקבל p -value שנסמן p_j .
 - נקבע סף כלשהו α .
 - נוסיף למודל את המשתנה x_j שעבורו p_j הוא הנמוך ביותר שעבורו גם $p_j < \alpha$.
 - נחזור על הפעולה עד שלא נוסיף עוד משתנים למודל.

• אלגוריתם שני - בחירה אחורה

- נתחיל עם המודל הכי גדול שנרצה לקחת בחשבון. כלומר $Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon$.
- הערה: בסימון זה נניח כי p הוא אינדקס המכיל למשל גם אינטראקציות או כל משתנה מסביר אחר שהיינו רוצים לקחת.
- לכל $j = 1, \dots, p$ נבצע:
 - * נגדיר מודל בלי המשתנה X_j . כלומר $Y = \beta_0 + \sum_{\substack{i=1 \\ i \neq j}}^p \beta_i X_i + \epsilon$.
 - * נבצע T -test להשערה $H_0 : \beta_j = 0$ ונקבל p -value שנסמן p_j .
 - נקבע סף כלשהו α .
 - נסיר מהמודל את המשתנה x_j שעבורו p_j הוא הגבוה ביותר שעבורו גם $p_j > \alpha$.
 - נחזור על הפעולה עד שלא נסיר עוד משתנים מהמודל.

• אלגוריתם שלישי - תת-מודל אידאלי

- נגדיר את $S = \{X_1, \dots, X_p\}$ להיות המודל הכי גדול שנרצה לקחת בחשבון.
- הערה: בסימון זה נניח ללא הגבלת הכלליות כי p הוא אינדקס המכיל למשל גם אינטראקציות או כל משתנה מסביר אחר שהיינו רוצים לקחת.
- לכל תת-קבוצה $A \subset S$ נתבונן במודל המתקבל מהמשתנים שב- A , כלומר $Y = \beta_0 + \sum_{X_j \in A} \beta_j X_j + \epsilon$.
- * בהתאם לקריטריון קבוע מראש אותו נסמן f , נחשב את $f(A)$.

- נגדיר את המודל להיות $S^* = \arg \max_{A \subset S} \{f(A)\}$.

הערה: נשים לב שמודל זה הוא כבד מבחינה חישובית, שכן מספר תתי הקבוצות של S שהיא מגודל p , הוא 2^p .

הערה: יש לדון איזה קריטריון f כדאי לקבוע לבדיקת טיב המודל. נדון בכך בפרק הבא.

17.1 מדידת טיב של מודל

מבוא: נתון מודל כלשהו $\hat{f}$. למשל המודל של רגרסיה לינארית הוא $\hat{f}(X_i) = X_i \hat{\beta}$.

כיצד נקבע עד כמה המודל הוא טוב?

ניתן להגדיר קריטריון על ידי $\bar{R}^2$ (כלומר $Adjusted R^2$). נציג קריטריון אחר שיתייחס למידת יכולתו של המודל לחזות נכון עבור תצפיות עתידיות.

הגדרה: יהי מדגם $(X_1, Y_1), \dots, (X_n, Y_n)$ כאשר $X_i \in \mathbb{R}^p$, ונניח כי הנתונים מקיימים $Y = X\beta + \epsilon$ עבור $\beta \in \mathbb{R}^p$ ועבור $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ ב"ת. יהי $\hat{f}$ מודל לינארי.

נגדיר את Mallows's C_p להיות:

$$C_p(\hat{f}) := \frac{1}{n} SSE(\hat{f}) + \frac{2p}{n}$$

נתייחס לזה כאל מדד לטיב המודל, כלומר נוכל לקבוע מודל לפי הקריטריון:

$$\hat{f}^* := \arg \min_{\hat{f}} \{C_p(\hat{f})\}$$

כעת נגדיר סטטיסטי חדש שבאמצעותו נוכל להראות מדוע C_p מהווה קריטריון לטיב המודל ביחס ליכולתו לחזות נכון עבור תצפיות עתידיות.

17.1.1 שגיאת התחזיות הצפויה (Expected Prediction Error)

הגדרה: יהי מדגם $(X_1, Y_1), \dots, (X_n, Y_n)$ כאשר $X_i \in \mathbb{R}^p$, ונניח כי הנתונים מקיימים $Y = X\beta + \epsilon$ עבור $\beta \in \mathbb{R}^p$ ועבור $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ ב"ת. יהי $\hat{f}$ מודל לינארי כלשהו.

נגדיר מדגם חדש $(X_1, \tilde{Y}_1), \dots, (X_n, \tilde{Y}_n)$ על ידי הגרלת $\tilde{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 I)$ ב"ת והגדרת $\tilde{Y}_i = \beta X_i + \tilde{\epsilon}_i$ (נשים לב כי רק ערכי Y חדשים במדגם זה).

נגדיר את שגיאת התחזיות הצפויה (Expected Prediction Error) על ידי:

$$EPE(\hat{f}) := \frac{1}{n} \sum_{i=1}^n E_{\tilde{Y}} \left[(\hat{f}(X_i) - \tilde{Y}_i)^2 \right]$$

⁹נחשוב על כך שלכל תת-קבוצה A , כל איבר $X_j \in S$ יש לו אחת משתי אפשרויות: להיות ב- A או לא להיות ב- A . יש p איברים, ולכן יש 2^p תתי-קבוצות.

הערה: באופן כללי עבור המודל $\hat{f}$ מוגדר גם סכום ריבועי השגיאות $SSE(\hat{f})$. נשים לב כי $SSE(\hat{f})$ מהווה מדד לשגיאות שמוטה לטובת המדגם הספציפי שנבחר (X_i, Y_i) , שכן משתמשים בו פעמיים - הן לצורך קביעת המודל והן לצורך מדידת השגיאות. לכן המדד $EPE(\hat{f})$, שאינו סובל מההטיה הזו, מהווה מדד עדיף על פני $SSE(\hat{f})$.

טענה: (ללא הוכחה) תחת תנאי ההגדרה לעיל, מתקיים תמיד:

$$\frac{1}{n} E_Y [SSE(\hat{f})] = E_Y [EPE(\hat{f})] - \frac{2p}{n}$$

$$E_Y [SSE(\hat{f})] < E_Y [EPE(\hat{f})] \text{ בפרט מתקיים תמיד}$$

הערה: האיבר $\frac{2p}{n}$ מכונה "Optimism of SSE ", שכן הוא מודד עד כמה $SSE(\hat{f})$ מהווה מדד "אופטימי" מידי של השגיאות (בגלל ההטיה לטובת המדגם המקורי).

מסקנה: $C_p(\hat{f})$ מהווה אומדן חסר הטיה של $EPE(\hat{f})$, שכן מתקיים:

$$E[C_p(\hat{f})] = E\left[\frac{1}{n} SSE(\hat{f}) + \frac{2p}{n}\right] = \frac{1}{n} E[SSE(\hat{f})] + \frac{2p}{n} = E[EPE(\hat{f})]$$

כאשר השוויון האחרון נובע מהטענה האחרונה.

הערה: Mallows's C_p מהווה מקרה פרטי של AIC - Akaike Information Criterion, המוגדר על ידי:

$$AIC(\hat{f}) := \sum_{i=1}^n \log \text{Likelihood}(Y_i) - \frac{2}{n} \# \{ \text{Parameters of the model } \hat{f} \}$$

מדד נוסף לטיב מודל הוא BIC - Bayesian Information Criterion, שפותח על ידי פרופ' גדעון שוורץ מהאוניברסיטה העברית.