רגרסיה מסוג Logit

גם המשתנה <u>המוסבר</u> שלי יכול להיות משתנה דמי. כאשר המשתנה המוסבר הוא משתנה איכותי, דיכוטומי נשתנש ברגרסיה מסוג logit.

<u>דוגמא</u>

- המשתנה המוסבר: האם הפרט לקה או לא לקה בהתקף לב ← משתנה איכותני.
- משתנה מסביר: משקל ← משתנה כמותי אינטרוולי.

נגדיר משתנה דמי D:

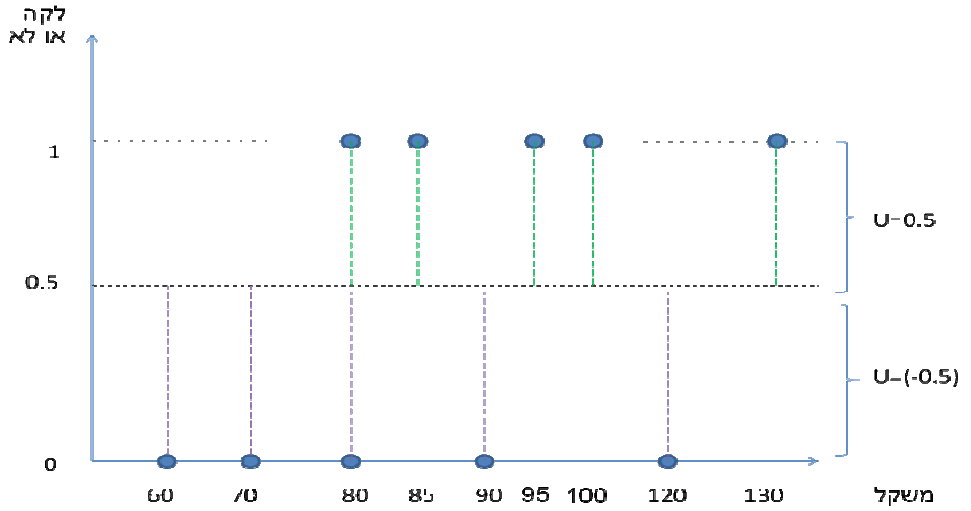
- מקבל ערך "1" אם לקה בהתקף לב.
- מקבל ערך "0" אם לא לקה בהתקף לב.

נתונים שנאספו:

מספר דגימה	X- משקל	התקף לב
1	60	0
2	100	1
3	70	0
4	85	1
5	90	0
6	95	1
7	80	0
8	120	0
9	80	1
10	130	10

יש 5 שלקו בהתקף לב ו-5 שלא לקו בהתקף לב. יש לנו 2 פרטים במשקל 80- אחד לקה ואחד לא. הסיבה היא משתנים מסבירים נוספים שיכולים להסביר את התקף הלב.

<u>דיאגראמת הפיזור</u>



© Tiano Yael-הסיכומים לא נבדקו ע"י המרצה

אני מעוניינת למצוא קו רגרסיה שיתאר לי את ההתפלגויות האלה (תצפיות). ניתן לראות כי במשקלים נמוכים הרוב נמצא ב-0 ובמשקלים הגבוהים הרוב נמצאים ב-1.

איזה קו רגרסיה נעביר על מנת שיתאר בצורה הטובה ביותר את דיאגראמת הפיזור?
(יש לי 5 סטיות שהם 0.5 ו-5 סטיות שהן (0.5-) ולכן סכום הסטיות הוא אפס.

<u>הנחה קלאסית 1:</u>
אין קשר בין הסטייה (U) לבין המשתנה המסביר (X), כלומר:

$$COV\ (Ui, Xi) = 0$$

ובדוגמא שלנו יש קשר כי במשקל נמוך הסטיות הן שליליות ובמשקל גבוה הסטיות הן חיוביות לכן יש קשר. ההנחה הקלאסית הזו לא מתקיימת. כי יש קשר בין ה-Xים לבין הסטיות.

<u>הנחה קלאסית 2:</u>
אין קשר בין סטייה אחת לסטייה שנייה.

$$COV\ (Ui, Uj) = 0$$

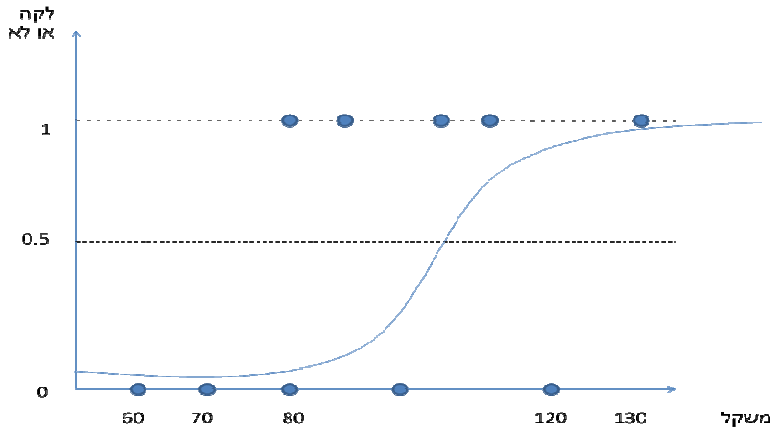
בדוגמא שלנו יש קשר בין הסטיות ולכן ההנחה הקלאסית לא מתקיימת. למשל: 0.5, 0.5, 0.5, ... , 0.5-, 0.5-, 0.5-. בגדול רואים פה חלוקה של שתי קבוצות.

משפט גאוס מרקוב אומר שהאומדנים בשיטת הריבועים הפחותים יהיו היעילים ביותר בתנאי שההנחות הקלאסיות מתקיימות. ופה אנו רואים שלא כל ההנחות הקלאסיות מתקיימות. לכן נצטרך לחפש שיטה אחרת.

לרגרסיה באופן כללי 2 מטרות:
1. הסבר (אני רוצה לדעת מה ההסתברות ללקות בהתקף לב בכל רמה ורמה של משקל).
2. ניבוי.

ולכן נשתמש ברגרסיה מסוג Logit.

רגרסיה מסוג Logit



© Tiano Yael-הסיכומים לא נבדקו ע"י המרצה

- ככל שהמשקל גדול יותר ההסתברות ללקות בהתקף לב הולכת וגדלה.
- תוספת של 10 ק"ג במשקל נמוך מעלה *מאוד* את ההסתברות ללקות בהתקף יותר מאשר תוספת של 10 ק"ג במשקל גבוה שהוא פחות משמעותית.
- במשקלים נמוכים הוא שואף לאפס.
- במשקלים נמוכים הוא שואף לאחד.
- הרגרסיה הזו היא לא לינארית- הקצב של השיפוע הולך וקטן.

Ln[(p/(1-p)] = a + bx + u

- p- הסתברות ללקות בהתקף לב.
- x- משקל.

<u>שאלה</u>
אם משקל האדם עולה (אגף ימין במשוואה עולה) כלומר X עולה אז אגף שמאל גם חייב לעלות. כלומר גם Ln[(p/(1-p)] יהיה חייב לעלות.

כאשר X **עולה** ← Ln[(p/(1-p)] **עולה** ← P **עולה**.

$\mathbf{Ln_e\ a = b}$
$\mathbf{e^b = a}$

אנחנו מניחים ש-u=0 כי זאת תוחלתו וכי a+ bx= 2.

$Ln[(p/(1-p)]= 2 \rightarrow e^2 = (p/(1-p) \rightarrow e^2 - p*e^2 = p$

$e^2 = p*(1 + e^2) \rightarrow \mathbf{p= e^2 / (1+e^2)} \rightarrow p=0.88$

<u>נוסחת קיצור</u>

$$P=\frac{e^{\,a+bx}}{1+e^{\,a+bx}}$$

לסיכום
רגרסיה מסוג Logit תשמש אותנו כאשר המשתנה המוסבר שלי איכותני (ערכים של 1 או 0). רגרסיה OLS לא טובה למשתנים איכותניים, כיוון שהנחות היסוד אינן מתקיימות.

מדוע סמנו ב-1 ללקות וב-0 לא ללקות ולא להפך?
זאת כיוון שמחשבים את ההסתברות שיקרה משהו (יקבל התקף לב) ולא את ההסתברות שלא יקרה משהו (לא לקה בהתקף לב). לכן, תמיד נסמן ב-1 עבור משהו שיקרה!

© Tiano Yael-הסיכומים לא נבדקו ע"י המרצה

מה קורה כאשר יש יותר מאפשרות אחת?
לדוגמא, האופציות הן:
1. טוב מאוד
2. טוב
3. לא טוב
4. גרוע

אם אני בודקת את ההסתברות להיות בריא נסמן:

"בריא"- 1 ← טוב מאוד + טוב

"לא בריא"- 0 ← לא טוב + גרוע

ככה הפכנו את זה לשימוש ברגרסיה מסוג Logit. אם היינו רוצים לבדוק את האפשרות שהוא חולה, היינו מסמנים את "לא בריא" ב-1 ואל "בריא" ב-0.

מה קורה אם יש לי 3 קטגוריות כתשובה?
כדי להריץ רגרסיה מסוג logit נחלק ל-2 קטגוריות באופן הבא:
- נאחד את קטגוריה 1 ו-2 לעומת 3 ונריץ רגרסיה.
- אח"כ נאחד את קטגוריה 1 לעומת 2 ו-3 ונריץ רגרסיה.

אחרי זה אני אחליט באיזו מהרגרסיות להשתמש. נבדוק איפה <u>רמת המובהקות</u> גבוהה יותר ואיפה <u>מקדם ההסבר</u> גבוה יותר. כך נבחר ברגרסיה המתאימה יותר.

<u>דוגמא</u>
חוקר מעוניין לבחון את ההסתברות להיות בעל דירה. הוא אסף את הנתונים הבאים: האם הפרט הוא בעל דירה, גיל, מספר שנות נישואין, רמת השכלה, רמת הכנסה ואיזור מגורים. הוא הריץ רגרסיה וקיבל את הרגרסיה הבאה:

$$\text{Ln}\,[p/\,(1-p)] = 0.05 + 0.05*\text{AGE} + 0.2*\text{INCOME} + 0.1*\text{MARRIGE} - 0.025*\text{STUDENT} - 0.1*\text{TEL AVIV} + e$$

1. <u>האם כל עליה בשנה גיל מעלה את ההסתברות להיות בעל דירה ב-5%?</u>
כל עליה בשנה גיל לא מעלה את p ב-5% אלא היא מעלה את $\text{Ln}\,[p/\,(1-p)]$ ב-5% ולכן זה יהיה פחות.

2. <u>למי מהקבוצות הבאות יש את ההסתברות הגבוהה ביותר להיות בעל דירה?</u>
   א. תושבי ת"א
   ב. תשובי באר שבע
   ג. תושבי הגליל
   ד. לא ניתן לדעת

   התשובה היא לא ניתן לדעת, אמנם לתושבי ת"א יש הסתברות נמוכה יותר להיות בעל דירה, אבל עדין לא ניתן לדעת האם תושבי הגליל או באר שבע הם בעלי סיכוי גדול יותר להיות בעלי דירה.

3. <u>מה משותף למסבירים גיל, מספר שנות נשואים והכנסה?</u>
כולם מעלים את ההסתברות להיות בעלי דירה (b הוא חיובי) וגם שלושתם כמותיים.

הסיכומים לא נבדקו ע"י המרצה-Tiano Yael ©

4. <u>מה תורם יותר להסתברות להיות בעל דירה? תוספת של 20 שנות גיל או תוספת של 10 שנות נשואים?</u>
התשובה היא אותו דבר כיוון שאם נכפיל את גדלים אלו ב-b של כל מקדם בהתאמה נקבל את הגודל 1. כלומר, ההסתברות תעלה בפחות מ-100%.

5. <u>נתון כי פרט מסוים הוא בן 35, הכנסתו 6000 ₪, הוא נשוי שנתיים, תושב ת"א ולומד באוניברסיטת ת"א. מה ההסתברות שלו להיות בעל דירה?</u>

$$\text{Ln}\,[p/\,(1-p)] = 0.05 + 0.05*35 + 0.2*6 + 0.1*2 - 0.025*1 - 0.1*1 = 3.08$$

$$P = \frac{e^{3.08}}{1 + e^{3.08}}$$

p= 0.956 ← ההסתברות גבוה שהוא יהיה בעל דירה.

איך נדע שהתשובה שלנו שגויה? ← אם יוצא הסתברות שהיא גדולה מ-1 אז בוודאות זה טעות.

SPSS ← רגרסיה LOGIT

Analyze → Regression →Binary logistic

<u>בדיקת מובהקות</u>
- בדיקת מובהקות ע"י לוח ANOVA שימוש ב-F (כמו ברגרסיה לינארית).
- דוחים את השערת האפס כש: $B_1=B_2= \ldots =B_n= 0$

כלומר, אם דחיתי את ההשערה קיים לפחות משתנה אחד שהוא שונה מאפס ואז זה מובהק כי הוא אכן מסביר את הרגרסיה.

- OLS ← לוח F.
- Logit ← לוח χ^2.

כדי לדעת אם רגרסיה מובהקת אז לפחות משתנה מסביר אחד צריך להיות שונה מאפס (זה F). Sig אומר לנו מהי רמת המובהקות.

אם כל המשתנים שווים לאפס זה אומר שאף משתנה אחד לא מצליח להסביר את הרגרסיה ואז דוחים את ההשערה.

© Tiano Yael-הסיכומים לא נבדקו ע"י המרצה