פרק ה-1

מודל הרגרסיה הלוגיסטית.

נעזר בחומר משקפים של ד"ר נויה גלאי וד"ר ניצה ברקן

מודל רגרסיה לוגיסטית (logistic regression model)

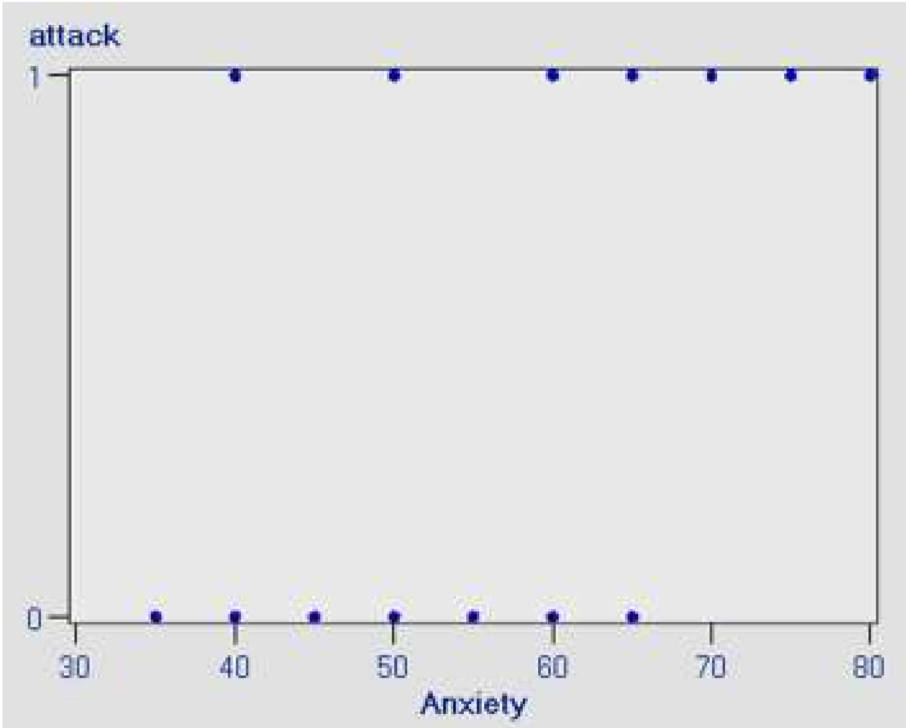
- מודל הרגרסיה הלוגיסטית מאפשר לתאר את ההסתברות למחלה תוך התייחסות בו זמנית למספר משתנים מסבירים.
- גישה זו היא יעילה ונוחה.
- המודל מציג את פונקצית Logit (הלוג'יט) של ההסתברות למחלה (תוצאה בינרית) כפונקציה לינארית של המשתנים המסבירים.

מודל רגרסיה לוגיסטית (logistic regression model)

- דוגמא: הקובץ heart_attack (ד"ר ניצה ברקן)
מעוניינים לבדוק האם קיים קשר בין מקרים של התקף לב לבין טיפול מסוים ורמת החרדה של מטופל.
משתנה תלוי: attack: 1= יקבל התקף לב, 0= לא יקבל התקף לב. משתנה בלתי תלוי:
anxiety (רמת החרדה): ככל שהערך גבוה, רמת החרדה גדלה.

מודל רגרסיה לוגיסטית (logistic regression model)

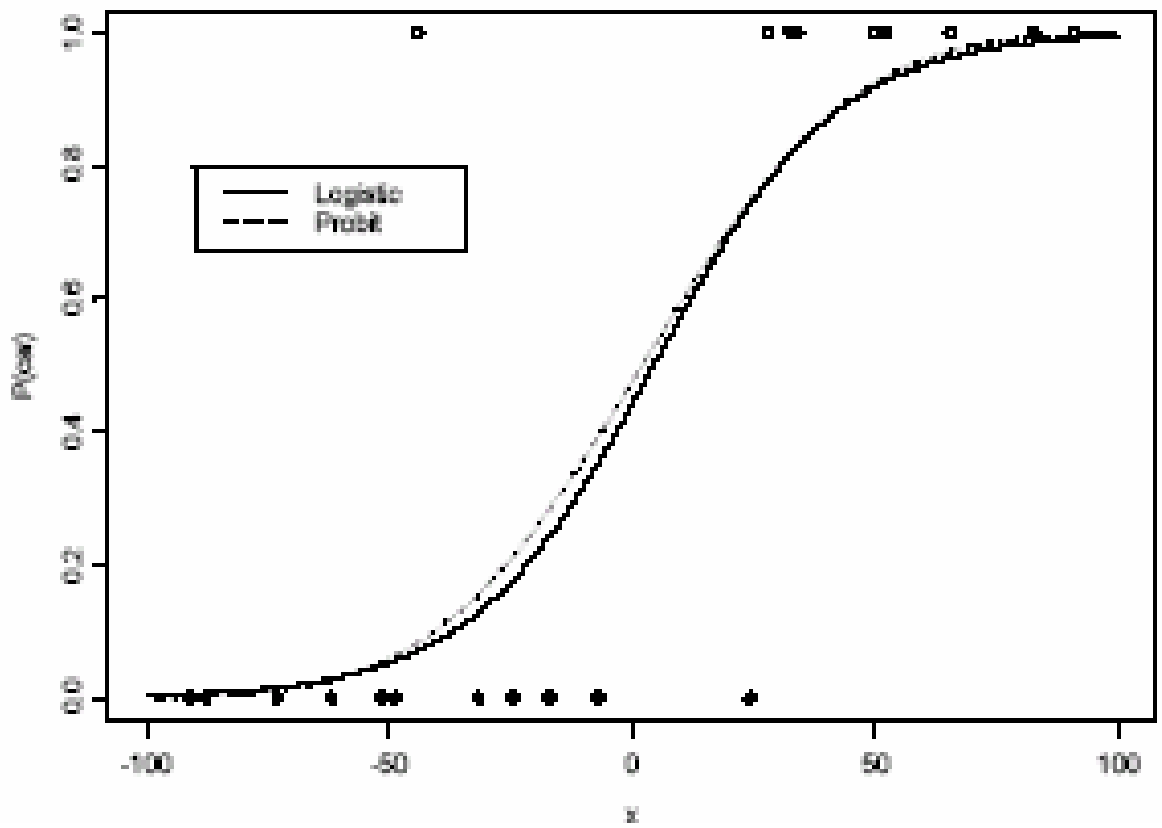
- הצגה גרפית של הנתונים: משתנה תלוי: attack
משתנה בלתי תלוי: anxiety



מודל רגרסיה לוגיסטית (logistic regression model)

- קיימת מגמה מסוימת עבור אותם האנשים שלא קיבלו התקף לב, ניתן לראות שרמת החרדה שלהם יותר קטנה מאשר אלה שכן קיבלו התקף לב.
- אבל הגרף הנה"ל אינו מספק תמונה ברורה לגבי מהות היחסים בין התקפי לב (attack) לבין רמת החרדה (anxiety).
- אחת השיטות להצגה ברורה יותר של הקשר בין התוצאה לבין משתנה בלתי תלוי, היא ליצור אינטרוולים עבור המשתנה הבלתי תלוי ולחשב ממוצעים של המשתנה התלוי בתוך כל אחד מהקטעים. הגרף המתקבל הינו מהצורה של אות *S*.

משמעות גרפית של המודל



הגדרת מודל הרגרסיה הלוגיסטית

- המודל המתאים לצורה כזו של עקומה הינו המודל הלוגיסטי (*logistic model*):

$$E(Y)=\pi(x)=\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

- נפעיל *logit transformation* ונקבל

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1 x$$

- כאשר הביטוי $\frac{\pi(x)}{1-\pi(x)}$ נקרא *odds* של $\pi(x)$.

אינטרפרטציה של הפרמטרים

- הגדרה: *odds ratio* מסומן ב-*OR* ומוגדר כיחס בין *odds* עבור *x=1* לבין *odds* עבור *x=0* כדלקמן

$$OR = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}$$

- טענה: עבור המשתנה המסביר הבינארי x,

$$OR = e^{\beta_1} \quad \Leftrightarrow \quad \ln(OR) = \beta_1$$

- הוכחה:

$$\ln(OR) = \ln\left(\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}\right) = \ln\left(\frac{\pi(1)}{(1-\pi(1)}\right) - \ln\left(\frac{\pi(0)}{(1-\pi(0)}\right) = \beta_1$$

אומדי נראות מכסימלית למקדמים ברגרסיה לוגיסטית

- עבור תוצאה בינארית שנסמנה באופן כללי על ידי

$$Y_i = \begin{cases} 1 \\ 0 \end{cases} \qquad E[Y_i] = \pi(\boldsymbol{x}_i)$$

- המודל

$$\ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} = \boldsymbol{x}_i^T \boldsymbol{\beta},\ i = 1,\ldots,n$$

- כאשר

$$\boldsymbol{x}_i = (1, x_{i1} \ldots x_{ip})^T; \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

- כך שמתקיים

$$E[Y_i] = \pi(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

אומדי נראות מכסימלית למקדמים ברגרסיה לוגיסטית

- האמידה היא בשיטת הנראות המכסימלית עם תוצאה המתפלגת ברנולית, בהינתן הערכים של המשתנים המסבירים. פונקצית הנראות המלאה היא:

$$L = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) P(\boldsymbol{X}_i = \boldsymbol{x}_i)$$

- כיוון שההתפלגות של גורמי הסיכון $\boldsymbol{X}_i$ אינה תלויה בפרמטרים של המודל, מספיק למצוא את המכסימום של

$$L^* = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$$

אומדי נראות מכסימלית למקדמים ברגרסיה לוגיסטית

- נמצא בדרך הרגילה את אמדי הנראות המכסימלית עבור הפרמטרים β , את האומדים לטעויות התקן דרך מטריצת האינפורמציה ומכאן מבחני מובהקות ורווחי סמך למקדמים. פונקצית הנראות הינה

$$L^*(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{x}_i^T \beta)}{1+\exp(\boldsymbol{x}_i^T \beta)} \right)^{y_i} \left(1 - \frac{\exp(\boldsymbol{x}_i^T \beta)}{1+\exp(\boldsymbol{x}_i^T \beta)} \right)^{1-y_i}$$

- לוג פונקצית הנראות:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\exp(\boldsymbol{x}_i{}'\beta)}{1+\exp(\boldsymbol{x}_i{}'\beta)} \right) + (1-y_i)\ln\left(1 - \frac{\exp(\boldsymbol{x}_i{}'\beta)}{1+\exp(\boldsymbol{x}_i{}'\beta)} \right) \right]$$

- את פונ' $l(\beta)$ נגזור לפי β ונקבל $p+1$ משוואות ה-*score*

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = 0$$

אומדי נראות מכסימלית למקדמים ברגרסיה לוגיסטית

- את מערכת משוואות הנ"ל לא ניתן לפתור אנליטית וניתן לחשב רק פתרונות נומריים כדי לקבל את אומדי הנראות המכסימלית $\hat{\beta}$ למקדמי הרגרסיה (למשל דרך שיטת Newton-Raphson).
- מטריצת האינפורמציה (Fisher Information Matrix) ממימד $(p+1)\times(p+1)$ מתקבלת על ידי הנגזרות החלקיות השניות של $l(\beta)$ ביחס ל- β והמטריצה ההופכית נותנת את מטריצת השונות של $\hat{\beta}$.

$$I(\beta)=E\left[-\frac{\partial^2 l(\beta)}{\partial\beta^2}\right] \qquad \Rightarrow \qquad Var(\hat{\beta})=I^{-1}(\beta)$$

אומדי נראות מכסימלית ומטריצת השונות

- דוגמא: הקובץ heart_attack
- Analysis ← regression ← logistic dependent=attack, מימין לשנות את ה- response variable ל-1 ← להעביר ל-quantitative variables את anxiety ← בלשונית statistics נסמן covariance matrix, על מנת לקבל את מטריצת השונויות ← finish

פלט SAS

Model Information		
Data Set	ECLIB000.HEART_ATTACK	
Response Variable	attack	attack
Number of Response Levels	2	
Number of Observations	20	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Response Profile

Ordered Value	attack	Total Frequency
1	0	10
2	1	10

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.0925	3.1710	5.0027	0.0253
Anxiety	1	0.1246	0.0553	5.0791	0.0242

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Anxiety	1.133	1.016	1.262

Estimated Covariance Matrix

Variable	Intercept	Anxiety
Intercept	10.05521	-0.17262
Anxiety	-0.17262	0.003058

אומדי נראות מכסימלית ומטריצת השונות

- <u>דוגמא:</u> הקובץ heart_attack
- Analysis ← regression ← logistic dependent=attack, מימין לשנות את ה- response variable ל-1 ← להעביר ל-quantitative variables את treat ואת anxiety ← בלשונית statistics נסמן covariance matrix, על מנת לקבל את מטריצת השונויות ← finish

פלט SAS

Model Information		
Data Set	ECLIB000.HEART_ATTACK	
Response Variable	attack	attack
Number of Response Levels	2	
Number of Observations	20	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Response Profile		
Ordered Value	attack	Total Frequency
1	0	10
2	1	10

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.3634	3.2139	3.9203	0.0477
Anxiety	1	0.1190	0.0550	4.6884	0.0304
treat	1	-1.0241	1.1711	0.7647	0.3818

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Anxiety	1.126	1.011	1.255
treat	0.359	0.036	3.565

Estimated Covariance Matrix			
Variable	Intercept	Anxiety	treat
Intercept	10.32914	-0.172	-0.4743
Anxiety	-0.172	0.003023	-0.0012
treat	-0.4743	-0.0012	1.371425