אקונומטריקה למדעי הנתונים

פרק 32 - רגרסיה לוגיסטית

תוכן העניינים

1. רגרסיה לוגיסטית 1

רגרסיה לוגיסטית:

רקע:

מתי נבצע רגרסיה לוגיסטית?

כאשר המשתנה המנובא הוא דיכוטומי (Binary Logistic):
יכול לקבל ערכים של 0 או 1.
הפונקציה הלוגיסטית מתארת את הסיכויים לקבל " 1" במשתנה התלוי כתלות במשתנים הב"ת.

הלוגיקה בניתוח רגרסיה לוגיסטית:

השוואת ניבוי Y ללא המשתנים המנבאים במודל לניבוי Y במודל הכולל את המשתנים המנבאים (סטטיסטי $\chi 2$).

טיב מודל הרגרסיה ("Goodness of fit"):

1. מובהקות המודל:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	12.225	4	.016
	Block	12.225	4	.016
	Model	12.225	4	.016

מבחן $\chi 2$ - תחת שורת ה-model נמצא את חי בריבוע ואת מובהקות המודל.

2. אחוז שונות מוסברת:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	96.524	.139	.189

$Nagelkerke\ R^2$ – מקביל ל-R^2 כללי ברגרסיה. אחוז שונות Y המוסברת ע"י כל המנבאים יחד (בטווח מוכר של 0-1).

3. דיוק בניבוי:

Classification Table[a]

Observed			Predicted: whether mom believes course will help: no	Predicted: whether mom believes course will help: yes	Percentage Correct
Step 1	whether mom believes course will help	no	46	5	90.2
		yes	17	14	45.2
	Overall Percentage				73.2

a. The cut value is .500

סגוליות (true negative) – ביחס ל- Y =0 במדגם, כמה המודל דייק בניבוי (90.2%).

רגישות (true positive) – ביחס ל- Y=1 במדגם, כמה המודל דייק בניבוי (45.2%).

אחוז הניבוי הכללי – בכמה בסה"כ המודל מדייק בניבוי (73.2%).

מושגים חשובים להבנת טבלת המקדמים:

ODDS:

"הסיכוי להתרחשות אירוע מסוים"- ההסתברות שהאירוע יקרה לעומת ההסתברות שאותו אירוע לא יקרה: $ODDS = \frac{p}{1-p}$

ODDS=1 – הסיכוי שהאירוע יתרחש שווה לסיכוי שהוא לא יתרחש ($\frac{0.5}{0.5}$).

ODDS>1 – הסיכוי שהאירוע יתרחש גבוה מהסיכוי שלא יתרחש (למשל- $\frac{0.75}{0.25}$).

ODDS<1 – הסיכוי שהאירוע יתרחש נמוך מהסיכוי שלא יתרחש (למשל- $\frac{0.25}{0.75}$).

ODDS RATIO (OR):

יחס בין סיכויים - $OR = \frac{ODDS_{(A)}}{ODDS_{(B)}}$

כיצד משתנה ההסתברות במעבר מקבוצה A לקבוצה B.

OR=1 – הסיכוי להתרחשות האירוע שווה בין שתי הקבוצות- אין קשר בין המב"ת למ"ת.

OR>1 – הסיכוי להתרחשות האירוע בקבוצה A גבוה מאשר בקבוצה B – קשר חיובי.

OR<1 – הסיכוי להתרחשות האירוע בקבוצה A נמוך מאשר בקבוצה B – קשר שלילי.

טבלת המקדמים – תרומות ייחודיות של כל מנבא:

(מקביל לטבלת Coefficients ברגרסיות לינאריות)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	EDU_YRS	-.107	.138	.603	1	.438	.898
	AGE	-.029	.020	2.078	1	.149	.971
	SATISFAC	.118	.175	.457	1	.499	1.126
	BIRTH#	.882	.321	7.530	1	.006	2.415
	Constant	.001	1.796	.000	1	.999	1.001

a. Variable(s) entered on step 1: EDU_YRS, AGE, SATISFAC, BIRTH#.

1. מבחן WALD למובהקות המשתנים:
מבטא את מובהקות המשתנה מבחינת תרומתו הייחודית לניבוי Y.

2. B – מקדמי המשתנים ב-log odds:
בטא חיובית – עלייה ב-log odds של Y כפונקציה של עליה ביחידה אחת של X.
בטא שלילית – ירידה ב-log odds של Y כפונקציה של עליה ביחידה אחת של X.

3. משוואת הרגרסיה:

$$\log\left(\frac{p}{1-p}\right) = \hat{\alpha} + \widehat{\beta_1}\, x_{1i} + \widehat{\beta_2} x_{2i} + \widehat{\beta_3} x_{3i} + \widehat{\beta_4} x_{4i}$$

חישוב הניבוי במונחי הסתברות (p): $p = \frac{1}{1+e^{-\log odds}}$

חישוב הניבוי במונחי סיכויים (ODDS): $\text{ODDS} = e^{\log odds}$

4. Exp (B) - יחס הסיכויים (Odds Ratio):
מבטא את <u>העלייה</u> (אם גדול מ-1) או את <u>הירידה</u> (אם קטן מ-1) בסיכויים להיות בעלי ערך '1' ב-Y כאשר הערך במשתנה המנבא גדל ביחידה אחת.

5. <u>היחס בין B ל-Exp (B)</u>: $\log \text{Exp(B)} = \text{B}$; $e^{B} = \text{Exp(B)}$

שאלות:

1) חוקרת בחוג למגדר ביקשה לבדוק האם מגדר משפיע על תעסוקה.
היא התבססה על סקר של הלמ"ס שדגם 826 מבוגרים בגילאי העבודה המרכזיים (25-55).
היא הגדירה את המשתנים באופן הבא:
WOMEN - "1" = אישה ; "0" = גבר.
WORKING - "1"= כן ; "0" = לא.
מהצלבה של שני המשתנים התקבלה הטבלה הבאה:

working * women Crosstabulation
Count

		women		Total
		.00	1.00	
working	.00	13	130	143
	1.00	338	345	683
Total		351	475	826

על סמך הטבלה חשבו:
א. מה ההסתברות של אישה לעבוד?
ב. מה הסיכוי של אישה לעבוד?
ג. מה ההסתברות של גבר לעבוד?
ד. מה הסיכוי של גבר לעבוד?
ה. מה יחס הסיכויים (OR) של נשים לעבוד לעומת גברים?
ו. מה הלוגריתם של יחס הסיכויים?
ז. מה יהיה ערך מקדם השיפוע B ברגרסיה הלוגיסטית לניבוי תעסוקה על פי מגדר ומה משמעותו?
ח. מה יהיה ערך Exp(B) ברגרסיה הלוגיסטית ומה משמעותו?

2) במחקר ביקשו לבדוק כיצד מצב משפחתי וגובה המשכורת משפיעים על בעלות על דירה.

משתני המחקר:

apartm - בעלות על דירה: "1"- כן ; "0"- לא.

status - מצב משפחתי: status (0) - רווק ; status (1) - בזוגיות ; status (2) - בזוגיות עם ילדים ; status(3) - פרוד או גרוש.

incom - הכנסה (<u>בעשרות</u> אלפי שקלים).

התקבלו הממצאים הבאים:

Classification Table[a]

			Predicted		
			apartm		Percentage Correct
	Observed		.00	1.00	
Step 1	apartm	.00	22	11	66.7
		1.00	10	22	68.8
	Overall Percentage				67.7

a. The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10.218	4	.037
	Block	10.218	4	.037
	Model	10.218	4	.037

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	79.876[a]	.145	.194

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Status			.682	3	.877	
	Status(1)	-.498	.713	.487	1	.485	.608
	Status(2)	-.520	.784	.441	1	.507	.594
	Status(3)	-.180	.748	.058	1	.810	.835
	income	.000	.000	8.580	1	.003	2.536
	Constant	-2.734	1.079	6.417	1	.011	.065

a. Variable(s) entered on step 1: Status, income.

א. האם ניתן לדחות את השערת האפס הטוענת כי אין קשר בין בעלות על דירה להכנסה ולסטטוס משפחתי?
ב. כמה אחוזים מצליחים המשתנים הב״ת להסביר מהשונות של המשתנה ״בעלות על דירה״?
ג. באיזה אחוז מצליח המודל לנבא באופן מדויק בעלות על דירה מתוך כלל המקרים?
ד. באיזה מידה מצליח המודל לנבא בהצלחה בעלות על דירה מתוך בעלי הדירה במדגם? כיצד נקרא המדד המתאים?
ה. באיזה מידה מצליח המודל לנבא בהצלחה אי-בעלות על דירה מתוך אלו שאינם בעלי דירה במדגם? כיצד נקרא המדד המתאים?
ו. מהי המשוואה לניבוי בעלות על דירה על סמך המשתנים הב״ת?
ז. לאיזה מהמשתנים הב״ת יש תרומה ייחודית מובהקת לניבוי בעלות על דירה? מהי משמעות מקדם B ו-Exp(B) של משתנה זה?
ח. על כל עליה ב-10,000 ₪ בהכנסה, בכמה אחוזים יעלה הסיכוי לבעלות על דירה?
   i. 53.6%.
   ii. 253.6%.
   iii. 153.6%.
   iv. 93%.
ט. על כל עליה של 20,000 ₪ בהכנסה, בכמה אחוזים יעלה הסיכוי לבעלות על דירה?
   i. 307%.
   ii. 423%.
   iii. 542%.
   iv. 642%.
י. מה ההסתברות של רווק המשתכר 20,000 ₪ להיות בעלים של דירה?
יא. האם ההסתברות של אותו רווק להיות בעל דירה גבוהה / שווה / קטנה מההסתברות שלו לא להיות בעל דירה?
יב. מהם הסיכויים (ODDS) שלו להיות בעל דירה?
יג. עבור איזה משכורת הסיכוי (הסתברות) של רווק להיות בעל דירה עולה על הסיכוי שלו לא להיות בעל דירה?
יד. במידה ומשתנה ההכנסה היה נמדד באלפי שקלים (ולא בעשרות אלפי שקלים), כיצד הדבר היה משפיע על ההשפעה השולית של מקדם ההכנסה, אם בכלל?

3) חוקרים בחנו את המאפיינים שעשויים לנבא את הביצוע של חניכים במבחן הסיום של קורס פקחי טיסה.
הביצוע במבחן נמדד על סולם של הצלחה/כשלון והמשתנים הבלתי תלויים כללו מין (1-זכר 0-נקבה), השכלה קודמת (0- ריאלית, 1- לא ריאלית) וביצוע במהלך הקורס (1-7).
להלן תוצאות ניתוח הרגרסיה:

Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	20.982	3	.000
	Block	20.982	3	.000
	Model	20.982	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	17.209[a]	.503	.699

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	מין	4.445	2.611	2.897	1	.089	85.161
	השכלה קודמת	-.146	2.054	.005	1	.943	.864
	ביצוע במהלך הקורס	2.283	.944	5.846	1	.016	9.810
	Constant	-19.284	8.056	5.731	1	.017	.000

a. Variable(s) entered on step 1: מין, השכלה, ביצוע במהלך הקורס

א. האם למודל הכולל את שלושת המנבאים יכולת הסבר משמעותית?
ב. כמה אחוזים מתוך השונות של Y מצליח המודל להסביר?
ג. מהי משוואת הניבוי?
ד. לאיזה מן המשתנים הב"ת תרומה מובהקת לניבוי?
ה. הסבירו את משמעות המקדמים (b) שהתקבלו עבור המשתנים הב"ת: מגדר, השכלה קודמת והביצוע במהלך הקורס.
ו. בטאו את המקדמים במונחי הסיכויים להצלחה בקורס (odds) והסבירו אותם.
ז. הועלתה הטענה כי ההסתברות ההצלחה של נשים בקורס היא נמוכה ביותר, גם אם הן בעלות השכלה ריאלית ושביצוען במהלך הקורס מקסימאלי. אנא בדקו את הטענה.
ח. עבור זכר, בעל השכלה ריאלית, מהי ההשפעה השולית של עליה ביחידה אחת בדירוג הביצוע במהלך הקורס על הסיכוי להצליח בקורס?

4) לפי מדגם של 20 זוגות נשואים, נאספו נתונים על המשתנה Y השווה ל-1 אם הזוג נוהג לצאת למסעדה לפחות פעם בשבוע ו-0 אחרת.

נאמד המודל: $p=\frac{1}{1+e^{-z}}$ כאשר $p=P(Y=1)$.

התקבלו התוצאות הבאות: $z=-9.456+0.368INCOM-1.207BABY$.

INCOM - ההכנסה של שני בני הזוג (באלפים). ההכנסה במדגם נעה בין 17 אלף ל- 44 אלף.

BABY - משתנה דמי המקבל את הערך '1' אם הזוג צריך להיעזר בשמרטפית ו-'0' אחרת.

ענה נכון/לא נכון:

א. זוג הנעזר בשמרטפית ומשתכר 30.5 אלף, יוצא למסעדה לפחות פעם בשבוע בהסתברות גבוהה מ-0.5.

ב. עבור זוג שאינו נעזר בשמרטפית, עליה של אלף שח בהכנסה, מעלה את ההסתברות לצאת למסעדה ב-0.368.

ג. כל אחד מערכי P הנאמדים כאן איננו גבוה יותר מ-0.99.

ד. הסיכוי של זוג, שהכנסתו עלתה ב-3000 ₪, לצאת למסעדה יעלה ב-200% בערך.

ה. המשכורת שצריך להרוויח זוג, אשר אינו נעזר בשמרטפית, כדי שהסיכוי שלו לצאת למסעדה יהיה שווה לסיכוי שלא לצאת למסעדה הוא 27,000.

ו. זוג, שלא נעזר בשמרטפית, צריך להרוויח יותר מ-28,000 ₪ כדי שהסיכוי שלו לצאת למסעדה יהיה גבוה פי 3 מהסיכוי שלו לא לצאת למסעדה.

ז. עבור odds ratio של משתנה "שמרטפות" התקבל רווח בר סמך הבא: $[0.123\ ;\ 1.01]$ ברמת ביטחון של 95%.

לפיכך ניתן לומר כי למשתנה "שמרטפות" תרומה מובהקת לניבוי הסיכוי לצאת למסעדה.

5) בשנה מסוימת הוגשו 750 בקשות לקבלת משכנתא ורק חלק מהן אושר.

המשתנה התלוי Y=1 אם הבקשה למשכנתא אושרה ול-0 אם נדחתה.

המנבאים:

S משתנה דמי השווה ל-1 אם מבקש המשכנתא הוא רווק ול-0 אחרת.

AGE = גיל בשנים.

המודל הנאמד הינו: $p=\frac{1}{1+e^{-z}}$ כאשר $p=P(Y=1)$.

$z=\alpha+\beta_1 age+\beta_2 age^2+\beta_3 S$.

תוצאות אמידת המודל: $z=-9.3+0.52age-0.006age^2-0.314S$

א. הסבירו את השפעת הגיל והמצב המשפחתי על ההסתברות לאישור המשכנתא.

ב. מה ההסתברות שתאושר משכנתא לרווק בן 30?

ג. עבור איזה גיל ההסתברות של אדם נשוי לקבל משכנתא היא מקסימאלית?

6) משרד הקבלה של האוניברסיטה רצה לבדוק באיזה מידה ניתן לחזות את ההצלחה של הסטודנט בקורס בסטטיסטיקה על סמך נתונים של מבחן פסיכומטרי, ציון ממוצע של תעודת בגרות וסוג תעודת הבגרות: ריאלית או לא ריאלית.
במדגם של 50 סטודנטים נאספו נתונים על המשתנה Y השווה ל-1 אם הסטודנט הצליח במבחן בסטטיסטיקה ו-0 אם נכשל.
כמו כן נרשמו עבור כל סטודנט ציון הפסיכומטרי, ממוצע הבגרות וסוג הבגרות (1 - בגרות ריאלית, 0 - לא ריאלית).
להלן התוצאות שהתקבלו:

	B	S.E.	Wald	df	Sig.
פסיכומטרי	.090	.046	3.723	1	.054
ציון בגרות	[redacted]	2.070	1.089	1	.297
בגרות_ריאלי	4.535	2.519	3.241	1	.072
Constant	-84.892	42.858	3.923	1	.048

א. באיזה שיטת ניתוח הייתם ממליצים להשתמש ומדוע?
ב. נתון כי ההסתברות להצליח בקורס בסטטיסטיקה עבור סטודנט שעשה בגרות הומנית, קיבל 690 בפסיכומטרי וציון 9 בבגרות הינה: 0.034.
ההסתברות של סטודנט שקיבל אותו ציון בפסיכומטרי, עם בגרות הומנית אבל ציונו בבגרות הוא 10 הינה: 0.233
על סמך הנתונים הללו השלם את הערך החסר בפלט המקדמים.
ג. לאיזה משתנים השפעה מובהקת על הסיכוי להצליח במבחן לסטטיסטיקה? (אלפא 10%)
ד. מה ההסתברות של סטודנט להצליח במבחן אם קיבל 680 בפסיכומטרי, ציון 10 בבגרות ולמד במגמה ריאלית?
ה. מהו השינוי בסיכויים (odds) להצליח במבחן בסטטיסטיקה כפונקציה של שינוי ביחידה אחת בפסיכומטרי?
ו. מהי ההשפעה השולית של נקודה נוספת בציון הבגרות על הסיכוי להצליח במבחן בסטטיסטיקה עבור סטודנט שקיבל 640 בפסיכומטרי ולמד במגמה ריאלית?
ז. רותי שיפרה את הפסיכומטרי שלה ב-20 נקודות.
בכמה יעלה הסיכוי שלה להצליח בקורס בסטטיסטיקה?
ח. אם החוקר היה מחליט לקודד בגרות שאינה ריאלית כ-1 ובגרות ריאלית כ-0, האם הדבר היה משפיע על ערכו של Exp(b) של סוג בגרות ועל המשמעות שלו?

תשובות סופיות:

1) א. 0.73 ב. 2.7 ג. 0.96 ד. 24 ה. 0.11

ו. 2.207- ז. $B=-2.207$. ח. $\text{Exp}(B)=0.11$.

2) א. כן. ב. 19.4% ג. 67.7% ד. רגישות = 68.8%

ה. סגוליות = 66.7%

ו. $In(odds)=-2.734-0.498status(1)-0.52status(2)-0.18status(3)+0.93\cdot incom$.

ז. משתנה "הכנסה". ח. 3. ט. 3. י. 0.3

יא. קטנה. יב. 0.42 יג. 29,400 יד. 0.093

3) א. כן. ב. 69.9% ג. $\log(odds)=-19.284+4.445x_{1i}-0.146x_{2i}+2.283x_{3i}$.

ד. המשתנה – "ביצוע במהלך הקורס". ה. ראו סרטון.

ו. מגדר - $\text{Exp}(b)=85.19$, השכלה קודמת - $\text{Exp}(b)=0.864$,

ביצוע במהלך הקורס - $\text{Exp}(b)=9.81$. ז. הטענה נכונה $(p=0.035)$.

ח. $\text{Exp}(b)=9.81$.

4) א. נכון. ב. לא נכון. ג. לא נכון. ד. נכון. ה. לא נכון.

ו. נכון. ז. לא נכון.

5) א. ראו סרטון. ב. 0.533 ג. 40.

6) א. רגרסיה לוגיסטית. ב. $B=2.16$. ג. "פסיכומטרי" ו-"בגרות ריאלית".

ד. 0.914 ה. 1.09 ו. 8.67 ז. 504% ח. ראו סרטון.