تحليل الانحدار اللوجيستي
Logistic Regression Model

1- الغرض من التحليل:

يعتبر نموذج الانحدار اللوجستي أحد النماذج الإحصائية التي تستخدم في كثير من المجالات التطبيقية كالمجالات الاقتصادية، والزراعية، والطبية، وغيرها من المجالات العلمية الأخرى، إذ يهتم هذا النموذج بدراسة وتحليل أثر عدة متغيرات مستقلة (لا يشترط نوعها) على متغير تابع نوعي مكون من مجموعتين أو أكثر من المجموعات المتنافية. ففي مثل هذه الحالات يصـــعب اســـتخدام نموذج الانحدار الخطي،لما تتطلبه من افتراضات غير محققة في الجانب التطبيقي.

2- الهيكل النظري لنموذج الانحدار اللوجستي الثنائي

بفرض أن المتغير التابع متغير نوعي ذو مجموعتين متنافيتين (اســـتحالة وقوعهما في آن واحد)، ويعبر عنه بمتغير ثنائي $y(0,1)$ ، فإن التوزيع الاحتمالي لهذا المتغير هو التوزيع البيرنوللي Bernoulli بمعلمة $\pi = pr(y=1)$ والتي تمثل احتمال النجاح ، وبفرض أن $(y_1, y_2, ..., y_n)$ تمثل المشاهدات لـ n من المتغيرات العشوائية المستقلة ، فإن نموذج الانحدار اللوجستي يعرف كالتالي:

$$\begin{aligned} y_i &\sim Bernoulli \quad (\pi_i), \\ \pi_i &= pr(y_i = 1 \mid X_{i1}, X_{i2}, ..., X_{ip}) \\ &= \frac{e^{\lambda_i}}{1+e^{\lambda_i}}, \quad i = 1,2,...,n \end{aligned} \qquad (1)$$

حيث تمثل $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ متجه معاملات الانحدار من الدرجة $((p+1)\times 1)$ ،وتمثل $(X_{i1}, ..., X_{ik})$ المشاهدات المفسرة للمفردة رقم i ، ويفترض أنه محدد Fixed ومعطاه ، وتعبر λ_i عن دالة اللوجت وتأخذ الصورة التالية.

$$\lambda_i = Ln(\pi_i / 1-\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, ..., \beta_p X_{ip} \qquad (1)$$

ويقوم برنامج (SPSS statistics, version 22,2013) بإيجاد تقدير لهذه المعاملات ويرمز لها بالرمز $\hat{\boldsymbol{\beta}}_{mle} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)'$

3- قاعدة التصنيف

تعرف النسبة $\pi_i/(1-\pi_i)$ بنسبة الرجحان (OR) Odds Ratio لحالة النجاح $(y_i = 1)$، وتقدير هذه النسبة هو :

$$OR_i = \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{\hat{\lambda}_i} \qquad (8)$$

ويلاحظ أن $OR_i > 0$ ، وعندما يكون إحتمال النجاح $\hat{\pi}_i$ أكبر من احتمال الفشل $(1-\hat{\pi}_i)$تكون نسبة الرجحان $OR_i > 1$ ، والعكس صحيح ، ومن ثم يمكن تصنيف المفردة لأحد مجموعتي المتغير التابع وفقا لقاعدة التصنيف التالية.

$$\begin{aligned} &The\ individual\ \ number\ \ i\ must\ \ be\ classified\ \ to\ success\ gr\ \ oup\ \ (y_i = 1)\ \ if\ \ OR_i > 1 \\ &The\ individual\ \ number\ \ i\ must\ \ be\ classified\ \ to\ failure\ gr\ \ oup\ \ (y_i = 0)\ \ if\ \ OR_i < 1 \end{aligned} \qquad (9)$$

وتدل القاعدة أعلاه أن المفردة رقم i يرجح تصنيفها لمجموعة النجاح $(y_i = 1)$ إذا كانت نسبة الرجحان أكبر من الواحد $(OR_i > 1)$، بينما يرجح تصنيفها لمجموعة الفشل $(y_i = 0)$ إذا كانت نسبة الرجحان أقل من الواحد $(OR_i < 1)$.

وبأخذ لوغاريثم نسبة الرجحان $Log_e(OR_i)$ يمكن التوصل إلى نموذج يأخذ شكل معادلة خطية في معاملات الانحدار اللوجستية يسمي بنموذج اللوجت Logit، ويأخذ الصورة التالية:

$$\begin{aligned} Logit\ (\hat{\pi}_i) &= Log_e(OR_i) \\ \hat{\lambda}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} \end{aligned} \qquad (10)$$

ويلاحظ من المعادلة أعلاه أن $Logit\ (\pi_i) > 0$ إذا كانت نسبة الرجحان $(OR_i > 1)$ ، وتكون $Logit\ (\pi_i) < 0$ إذا كانت نسبة الرجحان $(OR_i < 1)$ ومن ثم يمكن إعادة صياغة قاعدة التصنيف (9) من خلال نموذج اللوجت $\hat{\lambda}_i$ كما يلي:

$$\begin{aligned} &The\ individual\ \ number\ \ i\ must\ \ be\ classified\ \ to\ success\ gr\ \ oup\ \ (y_i = 1)\ \ if\ \ Logit\ \ (\pi_i) > 0 \\ &The\ individual\ \ number\ \ i\ must\ \ be\ classified\ \ to\ failure\ gr\ \ oup\ \ (y_i = 0)\ \ if\ \ Logit\ \ (\pi_i) < 0 \end{aligned} \qquad (11)$$

ومن ثم إذا كان المعامل $\beta_j > 0$ يحدث زيادة نسبية في نسبة الرجحان للنجاح $(y_i = 1)$، وإذا كان المعامل

$\beta_j < 0$ يحدث انخفاض نسبي في نسبة الرجحان للنجاح $(y_i = 1)$.

4- اختبارات الفروض.

بمجرد الحصـــول على تقديرات الإمكانية العظمى لمعاملات الانحدار اللوجســـتية يمكن إجراء العديد من الاختبارات الإحصـائية التي تسـعى لتحقيق الهدف من تطبيق نموذج الانحدار اللوجسـتي ومنها ما يلي:

اختبار مناسبة النموذج اللوجستي للتنبؤ.

يهدف هذا الاختبار إلى معرفة ما إذا كان نموذج الانحدار اللوجستي المقترح مناسب للتنبؤ بمجموعة المتغير التابع التي تنتمي لها المفردة ذات المشاهدات المفسرة $(x_1, x_2 ..., x_k)$. ومن ثم يأخذ فرض العدم H_0 والفرض البديل H_1 الصورة التالية:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \quad vs \quad H_1 : at \ least \ one \ of \ \beta_j \neq 0 \qquad (13)$$

ويستخدم اختبار نسبة الإمكان $Likelihood\ ratio\ (LR)$ لإجراء هذا الإختبار، حيث يعبر عن إحصائية الاختبار بالمعادلة التالية:

$$\chi^2 = -2 \log_e (LR) = -2[\mathrm{LL}(\hat{\boldsymbol{\beta}}_{mle} \mid H_0) - \mathrm{LL}(\hat{\boldsymbol{\beta}}_{mle} \mid H_1) \qquad (14)$$

حيث أن $\mathrm{LL}(\hat{\boldsymbol{\beta}}_{mle} \mid H_0)$ ، $\mathrm{LL}(\hat{\boldsymbol{\beta}}_{mle} \mid H_1)$ تعبر عن القيمة المحسوبة لـ $\mathrm{LL}(\boldsymbol{\beta})$ في حالة صحة الفرض العدم $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$، وفي حالة عدم صحته $H_1 : at \ least \ one \ of \ \beta_j \neq 0$. وتحت صحة الفرض العدم H_0 يقترب توزيع إحصائية الاختبار χ^2 من توزيع مربع كاي بدرجات حرية k أي أن $\chi^2 \approx \chi^2_{(k)}$.

اختبار معنوية تأثير المتغير المفسر.

يهدف هذا الإختبار إلى معرفة مدى الدلالة الإحصائية لتأثير كل متغير من المتغيرات المفسرة قيد الدراسة على المتغير التابع، ويعبر عن فرض العدم H_0 والفرض البديل H_1 كالتالي:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0 \ , \quad j = 1,2,..., k \qquad (15)$$

ويستخدم اختبار $(Wald)$ لإجراء هذا الإختبار، ويعبر عن إحصائية الاختبار بالمعادلة التالية:

$$wald = (\hat{\beta}_j / S.E_{\hat{\beta}_j})^2 \quad (16)$$

حيث أن $\hat{\beta}_j$ يعبر عن تقدير الإمكانية العظمى لمعامل الانحدار β_j والذي يعكس تأثير المتغير المفسر x_j على المتغير التابع، $S.E_{\hat{\beta}_j}$ يمثل الخطأ المعياري للتقدير $\hat{\beta}_j$. وتحت صحة الفرض العدم H_0 يقترب توزيع إحصائية الاختبار $wald$ من توزيع مربع كاي بدرجة حرية واحدة، أي أن $wald \approx \chi^2_{(1)}$.

اختبار جودة المطابقة.

ويقصد بهذا الاختبارمدى مطابقة القيم المتوقعة للمتغير التابع للقيم المشاهدة ، وفي هذا الاختبار يعبر عن فرض العدم H_0 والفرض البديل H_1 كالتالي:

$$H_0: \text{The expected values from using prediction model match the observed values.}$$
$$vs\ H_1: \text{The expected values from using prediction model does not match the observed values.} \quad (17)$$

ويدل الفرض العدم H_0 على أن القيم المتوقعة من استخدام نموذج التنبؤ تطابق القيم المشاهدة، وقبول هذا الفرض يستدل منه على أن النموذج تمثيل جيد للبيانات وأن ليس به قصور. وتستخدم طريقة Hosmer and Lemeshow(2000) لإجراء هذا الإختبار، حيث تستند هذه الطريقة على فكرة تقسيم المشاهدات الكلية n إلى عدد G من المجموعات المتساوية هي: $n_1, n_2, ..., n_G$ ، ثم ترتب احتمالات النجاح المتنبأ بها $\hat{\pi}_i's$ تصاعديا ويحسب متوسط احتمالات النجاح لكل مجموعة ففي المجموعة g يكون المتوسط هو $\bar{\pi}_g = \sum_{i=1}^{n_g} \hat{\pi}_i / n_g$ ، ومن ثم تحسب إحصائية الاختبار HL بالمعادلة التالية .

$$HL = \sum_{g=1}^{G} \left[\frac{(O_g - n_g \bar{\pi}_g)^2}{n_g \bar{\pi}_g (1 - \bar{\pi}_g)} \right] \quad (18)$$

حيث أن O_g هو التكرار المشاهد لعدد حالات النجاح داخل المجموعة g ، $\sum_{g=1}^{G} n_g = n$ ، وتحت صحة الفرض العدم H_0 يقترب توزيع الإحصائية HL من توزيع مربع كاي بدرجات حرية $(G-2)$ أي أن $HL \approx \chi^2_{(G-2)}$.

نسبة التصنيف الصحيح

يقصد بالتصنيف الصحيح للمفردة، أنها إذا كانت فعلا تنتمي للمجموعة $k(success, failure)$ وتم تصنيفها وفقا للمعيار (9) أو (11) لنفس المجموعة $k(success, failure)$ ، وعلى هذا الأساس يمكن

حساب نسبة التصنيف الصحيح لكل مجموعة وكذلك نسبة التصنيف الصحيح العام. وكلما كانت نسبة التصنيف الصحيح عالية كان ذلك مؤشرا لجودة النموذج.

تطبيق

المتغير التابع : موقف العائلة من تفضيلها أوعدم تفضيلها شراء سلعة معينة، حيث أن

$y = 1$: إذا كانت العائلة تفضل شراء السلعة. $y = 0$: إذا كانت العائلة لا تفضل شراء السلعة.

المتغيرات المفسرة:

incom : الدخل بالأف ريال *Family* : حجم الأسرة

اليانات

No	التفضيل	الدخل	حجم الأسرة
1	0	2.1	1
2	0	2.6	1
3	0	3.1	3
4	0	4.6	2
5	0	3.3	2
6	0	3.3	2
7	0	3.3	2
8	0	5.6	2
9	0	3.8	4
10	0	4.2	4
11	0	4.2	4
12	0	4.4	3
13	1	4.2	4
14	1	4.6	3
15	1	3.2	4

No	التفضيل	الدخل	حجم الأسرة
16	1	4.9	3
17	1	5	1
18	1	5	3
19	1	5	3
20	1	5.2	4
21	1	5.4	4
22	1	5.5	4
23	1	3.8	5
24	1	5.8	5
25	1	6.7	5
26	1	6.7	6
27	1	7.6	6
28	1	7.8	6
29	1	8.6	6
30	1	9.7	8

النتائج وتفسيرها

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	30	100.0
	Missing Cases	0	.0
	Total	30	100.0
Unselected Cases		0	.0
Total		30	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	19.028	2	.000
	Block	19.028	2	.000
	Model	19.028	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	21.352[a]	.470	.635

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

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.186	8	.327

Contingency Table for Hosmer and Lemeshow Test

		التفضيل = No		التفضيل = Yes		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2	1.990	0	.010	2
	2	3	2.836	0	.164	3
	3	1	2.370	2	.630	3

	4	3	1.683	0	1.317	3
	5	2	1.486	2	2.514	4
	6	0	1.140	4	2.860	4
	7	1	.420	2	2.580	3
	8	0	.074	3	2.926	3
	9	0	.002	3	2.998	3
	10	0	.000	2	2.000	2

Classification Table[a]

			Predicted		
			التفضيل		Percentage
	Observed		No	Yes	Correct
Step 1	التفضيل	No	9	3	75.0
		Yes	2	16	88.9
	Overall Percentage				83.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Incom	1.692	.835	4.110	1	.043	5.430
	Family	.963	.572	2.835	1	.092	2.620
	Constant	-10.359	4.694	4.871	1	.027	.000

Variables in the Equation

		95% C.I.for EXP(B)	
		Lower	Upper
Step 1[a]	Incom	1.058	27.878
	Family	.854	8.037
	Constant		

a. Variable(s) entered on step 1: Incom, Family.

```
Step number: 1

Observed Groups and Predicted Probabilities
```

```
       8 +
+
         I
I
         I
I
F        I
I
R      6 +
Y+
E        I
YI
Q        I
YI
U        I
YI
E      4 +
Y+
N        I
YI
C        I     N                                                      Y
YI
Y        I     N                                                      Y
YI
       2 +N    N                                                      N
Y                           Y+
         IN    N                                                      N
Y                           YI
         IN    N   N               Y  Y     N             NN       Y       N
YY YN               Y  YY   YYI
         IN    N   N               Y  Y     N             NN       Y       N
YY YN               Y  YY   YYI
Predicted ---------+---------+---------+---------+---------+---------+-----
----+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6
.7        .8        .9          1
  Group:
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYY
YYYYYYYYYYYYYYYYYYYYYYYYY

          Predicted Probability is of Membership for Yes
          The Cut Value is .50
          Symbols: N - No
                   Y - Yes
          Each Symbol Represents .5 Cases.
```

التفضيل	الدخل	حجم الأسرة	Predicted probability	Predicted group
0	2.1	1	0.00289	0
0	2.6	1	0.00671	0
0	3.1	3	0.09753	0
0	4.6	2	0.343	0
0	3.3	2	0.0547	0
0	3.3	2	0.0547	0
0	3.3	2	0.0547	0
0	5.6	2	0.73924	1
0	3.8	4	0.48061	0
0	4.2	4	0.64548	1
0	4.2	4	0.64548	1
0	4.4	3	0.49366	0
1	4.2	4	0.64548	1
1	4.6	3	0.57762	1
1	3.2	4	0.25109	0
1	4.9	3	0.69437	1
1	5	1	0.28167	0
1	5	3	0.72905	1
1	5	3	0.72905	1
1	5.2	4	0.90815	1
1	5.4	4	0.93274	1
1	5.5	4	0.94261	1
1	3.8	5	0.70794	1
1	5.8	5	0.9862	1
1	6.7	5	0.99696	1
1	6.7	6	0.99884	1
1	7.6	6	0.99975	1
1	7.8	6	0.99982	1
1	8.6	6	0.99995	1
1	9.7	8	1.0	1

التعليق على النتائج

- **يقوم الطالب بالتعليق من خلال ما تم فهمه في المحاضرة**