Logisztikus regresszió

9. előadás

Kvantitatív statisztikai módszerek

Dr. Szilágyi Roland

		Független változó (x)	
		Nem metrikus	Metrikus
Függő változó (y)	Nem metrikus	Kereszttábla elemzés	Diszkriminancia-analízis, **Logisztikus regresszió**
	Metrikus	Varianciaanalízis	Korreláció- és regresszióelemzés

Logisztikus regresszió előnyei:

Mind metrikus, mind nem metrikus független változók használatát megengedi

Kevesebb feltétel teljesülését kívánja meg

Logisztikus regresszió

- Olyan többváltozós módszer, amely segítségével esetek kategorizálását végezhetjük el a függő változó kategóriái szerint.
- Ellenőrizzük, hogy a csoporthoz való tartozás becsülhető-e, és ha igen, milyen arányban.
- Lehet:
  - **Kétváltozós (a függő változó bináris)**
  - Többváltozós

A logisztikus regresszió számítás célja

- A megfigyelési egységek valamely csoportba sorolása.
- A csoportosítás pontosságának mérése.
- Tehát beazonosítani azokat a tényezőket, amelyek szignifikánsan megkülönböztetik az esetek csoportjait, és ellenőrizni, hogy a csoporthoz való tartozás becsülhető-e adott független változókkal, és ha igen, hány százalékban.

Alkalmazási területei

- Piackutatás
- Vásárlási modellezés (vásárol - nem vásárol)
- Megbízhatóság (vissza fizeti a hitelt, vagy nem)
- Cégvizsgálat (csődös, vagy nem)
- Stb.

A vizsgálat menete

1. Cél, probléma megfogalmazása
2. Feltételek vizsgálata
3. Elemzés elvégzése
4. Eredmények értelmezése
5. Érvényesség értékelése

A logisztikus regresszó-számítás feltételei

1. A változók mérési szintje

- A függő változó nominális skálán mérhető, --lehet kétcsoportos ilyen esetben (binomiális 0/1),

  -lehet több csoportos is (multinomiális).
- A független változók bármilyen skálán mérhetők. (A nominális változók interakciói is szerepeltethetők a modellben).

A logisztikus regresszó-számítás feltételei

2. Az adatok függetlensége

Az összes megfigyelésnek függetlennek kell lennie egymástól, vagyis az adatok nem lehetnek korreláltak. Erre az adatgyűjtésnél, mintavételnél komoly figyelmet kell fordítani.

A logisztikus regresszó-számítás feltételei

3. A mintanagyság

Itt is kritikus pont a megfigyelések számának és a független változók számának aránya.

- Legalább 60 elemű minta szükséges.
- A teljes mintanagyság legalább tízszer nagyobb legyen a független változók számánál.

A logisztikus regresszó-számítás feltételei

4. Normalitás

A független változóknak normális eloszlásúnak kell lenniük. A feltétel sérülhet a kiugró értékek és a helytelen skálás miatt is.

- Egyváltozós normalitás tesztelése: boxplot, QQ ábra, hipotézis vizsgálat.
- Többváltozós normalitás tesztelése: Mahalanobis-mutató

A logisztikus regresszó-számítás feltételei

5. Multikollinearitás

Számos eddigi feltétel a többváltozós regresszió számításnál is megtalálható volt, hasonlóan a multikollinearitáshoz. A logisztikus regressziónál is feltételeznünk kell, hogy a független változók csak a függő változóval függnek össze, egymással nem.

Logisztikus regresszió

- „A logisztikus regresszió két egymást kölcsönösen kizáró kategória bekövetkezési esélyeinek az egymáshoz való arányát, vagyis az *odds* mértékét modellezi x_i magyarázó változók értékeinek az ismeretében.”

$$odds_x = \frac{P_x}{1 - P_x}$$

Forrás: Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Logisztikus regresszió

- „A logisztikus regresszió feltételezése szerint az odds logaritmusa – másképpen a siker valószínűségének logitja – a magyarázó változók lineáris függvénye.”

$$Y = \ln(odds_x) = \log it(P_x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

$$odds_x = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

Forrás: Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Logisztikus regresszió

$$odds_x = \frac{P_x}{1 - P_x}$$

$$P_x = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

Forrás: Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Maximum Likelihood módszer

A maximum likelihood módszer célja, hogy adott mérési értékekhez, az ismeretlen paramétereknek olyan becslését adja meg, amely mellett az adott érték a legnagyobb valószínűséggel következik be. Az eljárás a *likelihood függvény* maximalizálásával történik.

Forrás: Wikipedia

Maximum Likelihood módszer

- Az adott kimenet valószínűségét előrejelző függvény paramétereinek (β) becsült értékei adott x_i magyarázó változók mellett a Likelihood függvény maximumában találhatók, vagyis ahol:.”

$$L = \prod_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} \rightarrow \max$$

Forrás: Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Modell tesztelés

A modell illeszkedésének jóságát a **Hosmer-Lemeshow** teszt segítségével vizsgáljuk.

Ho: illeszkedik

H1: nem illeszkedik

Az egyedeket a becsült valószínűségek alapján rangsorba rendezi, majd valamely kvantilis (decilis) által meghatározott csoportokon χ^2 tesztet hajt végre.

β paraméterek tesztelése

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

$$\text{Wald}_i = \left(\frac{b_i}{s(b_i)} \right)^2$$

A modell magyarázóereje

- Reziduális négyzetösszegre alapozott mutató (lineáris regresszió)
- Likelihood arányra alapozott mutatók (az elkészített modell Likelihoodját egy alapmodelléhez viszonyítja)
- A helyes előrejelzések részaránya

Pseudo R^2

- A *Cox and Snell* R^2 a modell log likelihoodjának értékét egy alapmodell log likelihood értékéhez viszonyítja. A mutató elméleti maximális értéke (ami egy tökéletes modellt feltételez) kisebb, mint egy.
- A *Nagelkerke* R^2 az előző mutató skálázási problémáinak korrigálásával határozható meg.

Outputok

Classification Table[a,b]

			Predicted					
			Selected Cases[c]			Unselected Cases[d,e]		
			Previously defaulted		Percentage Correct	Previously defaulted		Percentage Correct
	Observed		No	Yes		No	Yes	
Step 0	Previously defaulted	No	375	0	100,0	142	0	100,0
		Yes	124	0	,0	59	0	,0
	Overall Percentage				75,2			70,6

a. Constant is included in the model.

b. The cut value is ,500

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	3,292	8	,915
2	11,866	8	,157
3	9,447	8	,306
4	4,027	8	,855

Forrás: Help- IBM SPSS Statistics

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	498,012[a]	,116	,172
2	447,301[b]	,201	,299
3	411,553[b]	,257	,381
4	394,721[c]	,281	,417

Forrás: Help- IBM SPSS Statistics

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c,d]		
			Previously defaulted		Percentage Correct	Previously defaulted		Percentage Correct
	Observed		No	Yes		No	Yes	
Step 1	Previously defaulted	No	361	14	96,3	137	5	96,5
		Yes	100	24	19,4	45	14	23,7
	Overall Percentage				77,2			75,1
Step 2	Previously defaulted	No	351	24	93,6	136	6	95,8
		Yes	80	44	35,5	36	23	39,0
	Overall Percentage				79,2			79,1
Step 3	Previously defaulted	No	348	27	92,8	135	7	95,1
		Yes	72	52	41,9	28	31	52,5
	Overall Percentage				80,2			82,6
Step 4	Previously defaulted	No	352	23	93,9	130	12	91,5
		Yes	67	57	46,0	27	32	54,2
	Overall Percentage				82,0			80,6

a. The cut value is .500

Classification table (Confusion matrix)

		előrejelzés (predicted)		
		no (0)	yes (1)	
valós állapot (observed)	no (0)	valós negatív (VN)	álpozitív (ÁP)	→specificitás VN/(VN+ÁP)
	yes (1)	álnegatív (ÁN)	valós pozitív (VP)	→ szenzitivitás VP/(ÁN+VP)
		↓ negatív prediktív érték VN/(VN+ÁN)	↓ pozitív prediktív érték relevancia/precizitás VP/(ÁP+VP)	pontosság (VP+VN)/ (VN+ÁP+ÁN+VP)

Előrejelzési képesség értelmezése

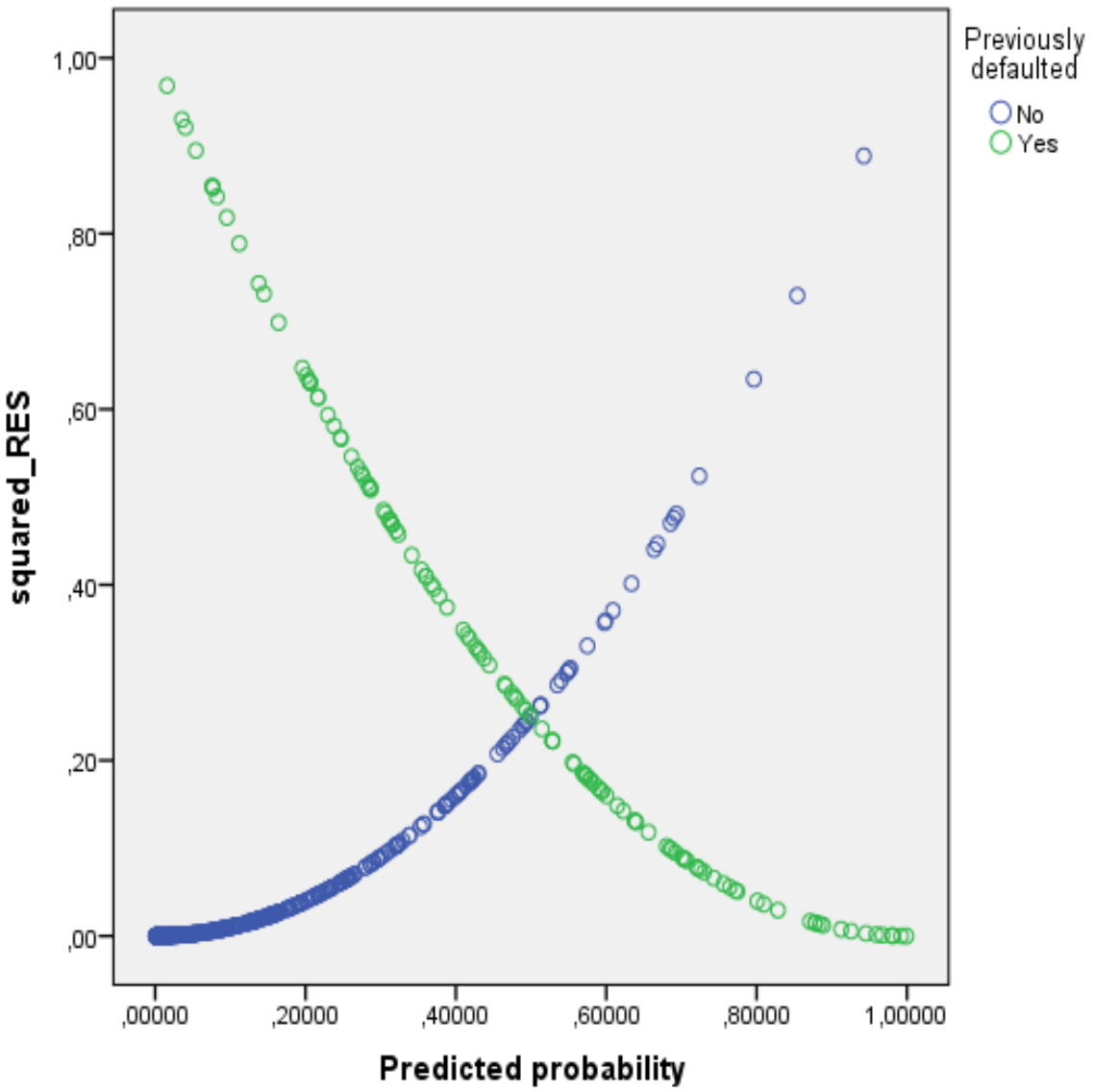
- Szenzitivitás: annak a valószínűsége, hogy az előrejelzés „1” lesz egy olyan ügyfél esetében, aki késik (default).
- Specificitás: annak a valószínűsége, hogy az előrejelzés „0” lesz egy olyan ügyfél esetében, aki nem késik.
- Pozitív prediktív érték: annak a valószínűsége, hogy az „1” előrejelzés esetében az ügyfél valóban késik.
- Negatív prediktív érték: annak a valószínűsége, hogy „0” előrejelzés esetében az ügyfél nem késik.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Debt to income ratio (x100)	,121	,017	52,676	1	,000	1,129	1,092	1,166
	Constant	-2,476	,230	116,315	1	,000	,084		
Step 2[b]	Years with current employer	-,140	,023	38,158	1	,000	,869	,831	,909
	Debt to income ratio (x100)	,134	,018	54,659	1	,000	1,143	1,103	1,185
	Constant	-1,621	,259	39,038	1	,000	,198		
Step 3[c]	Years with current employer	-,244	,033	54,676	1	,000	,783	,734	,836
	Debt to income ratio (x100)	,069	,022	9,809	1	,002	1,072	1,026	1,119
	Credit card debt in thousands	,506	,101	25,127	1	,000	1,658	1,361	2,021
	Constant	-1,058	,280	14,249	1	,000	,347		
Step 4[d]	Years with current employer	-,247	,034	51,826	1	,000	,781	,731	,836
	Years at current address	-,089	,023	15,109	1	,000	,915	,875	,957
	Debt to income ratio (x100)	,072	,023	10,040	1	,002	1,074	1,028	1,123
	Credit card debt in thousands	,602	,111	29,606	1	,000	1,826	1,470	2,269
	Constant	-,605	,301	4,034	1	,045	,546		

Paraméter értekmezés

- X_i 1 egységnyi növekedése esetén az „1”-es előrejelzés esélye átlagosan EXP(B) szeresére változik, minden egyéb változatlansága mellett.



Forrás: Help- IBM SPSS Statistics

Köszönöm a figyelmet

strolsz@uni-miskolc.hu