LOGIT-REGRESSZIÓ

a függő változó: névleges vagy sorrendi skála

a független változó: névleges vagy sorrendi vagy folytonos skála

Bevezetés – Mit jelent a logit regresszió?

Legyen a szakmai kérdés: Mi a valószínűsége, hogy adott életkorú (adott fizetéssel rendelkező) ember visszafizeti a hitlét?

Y: visszafizeti-e a hitelt
(igen-nem típusú változó)

x: fizetés (életkor)
(folytonos változó)

$$Y = \pi = f(x)$$

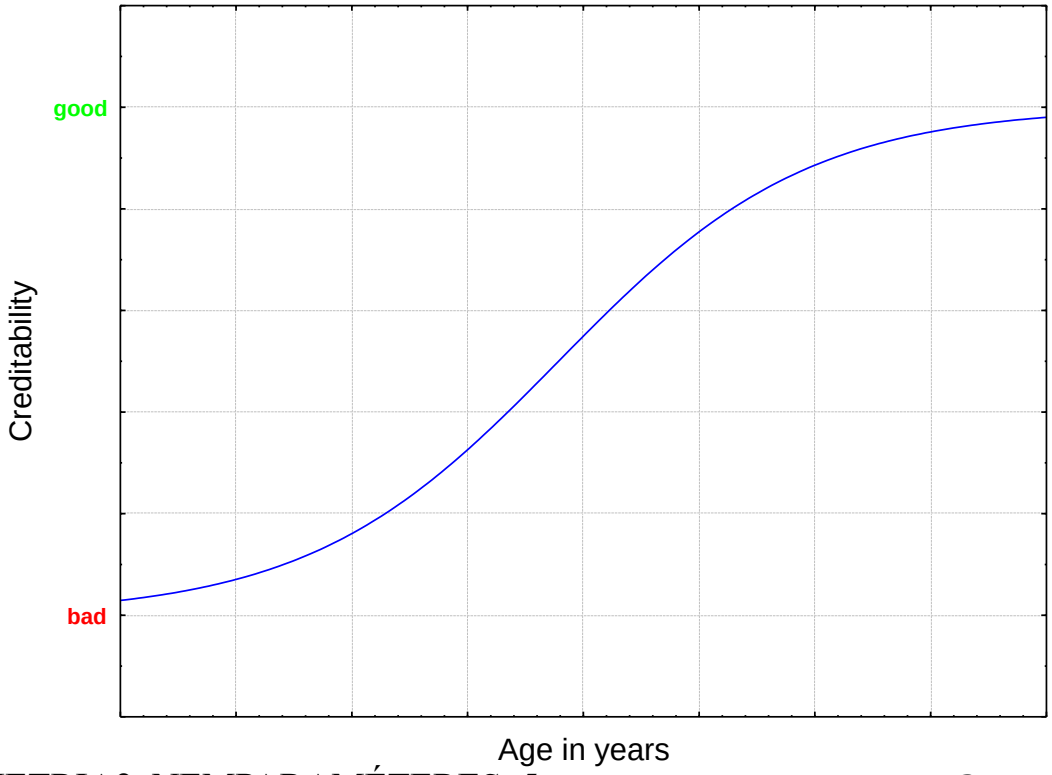
π: visszafizetés valószínűsége

$$0 \le \pi \le 1$$

általában nem lineáris fv.

↓

Transzformáljuk!

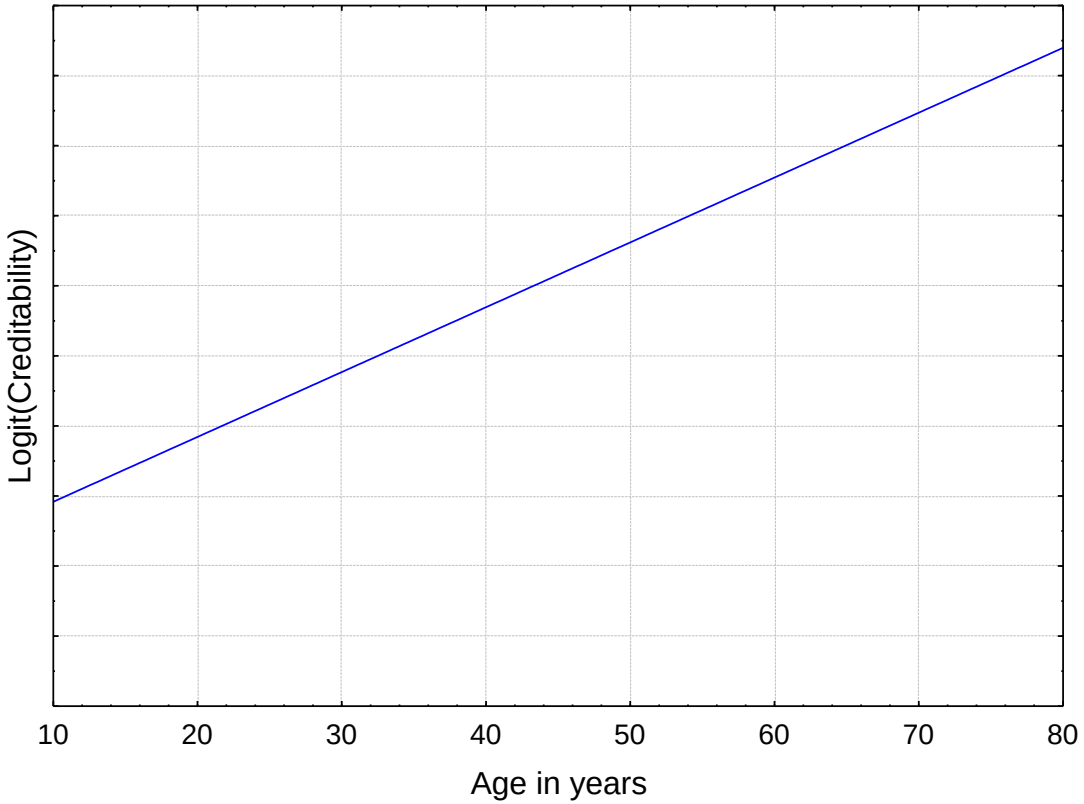


Transzformáció célja: linearizáljuk a függvényt $Y^{tr} = \alpha + \beta x$

$\pi \rightarrow$ esély $\rightarrow$ **ln(esély)=logit**

$(0,1) \quad (0, \infty) \quad (-\infty, \infty)$

$$\text{esély} = \frac{\pi}{1-\pi}$$



$$\text{logit} = \ln \frac{\pi}{1-\pi}$$

logit az esély logaritmusa

$$\text{logit} = \ln \frac{\pi}{1-\pi} = \alpha + \beta x$$

lineáris függvény

$$\text{logit} = \ln \frac{\pi}{1-\pi} = \alpha + \beta x$$

összefüggés regressziós módszerekkel kiszámítható:
azaz megadható, hogy hogyan függ a hitel visszafizetésének esélyének logaritmusa az életkortól

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

az így becsült logit értékből visszaszámolható a hitel visszafizetésének valószínűsége

Ahogy a szokásos regressziónál, itt is a következő feladataink vannak:

- ✓ a függvény paramétereinek (α és β) becslése
- ✓ a függvény alkalmasságának vizsgálata
- ✓ statisztikai próbák a függvényre vagy paramétereire
- ✓ konfidencia-tartományok számítása a paraméterekre

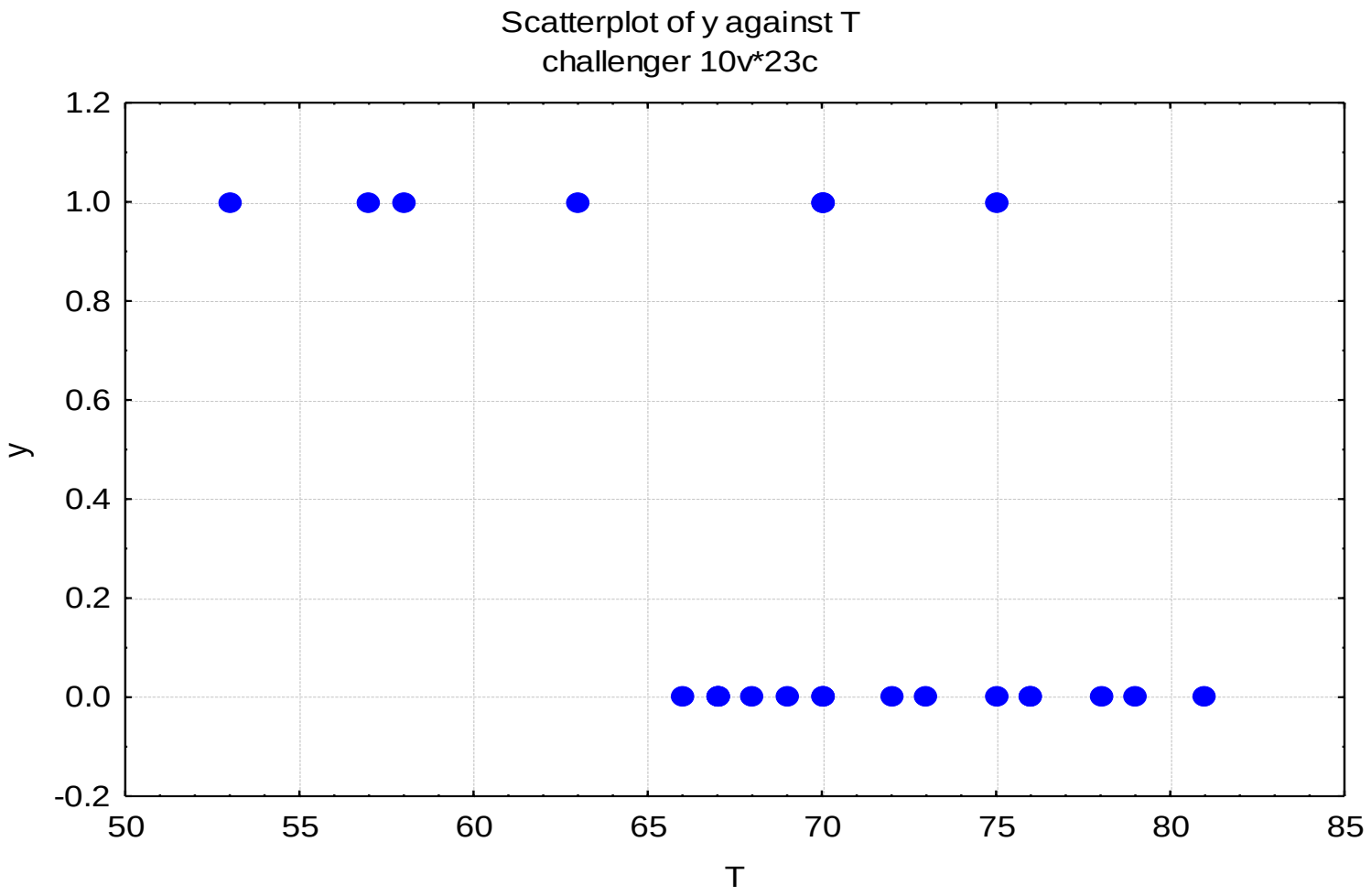
	2 T	3 y
1	66	0
2	70	1
3	69	0
4	68	0
5	67	0
6	72	0
7	73	0
8	70	0
9	57	1
10	63	1
11	70	1
12	78	0
13	67	0
14	53	1
15	67	0
16	75	0
17	70	0
18	81	0
19	76	0
20	79	0
21	75	1
22	76	0
23	58	1

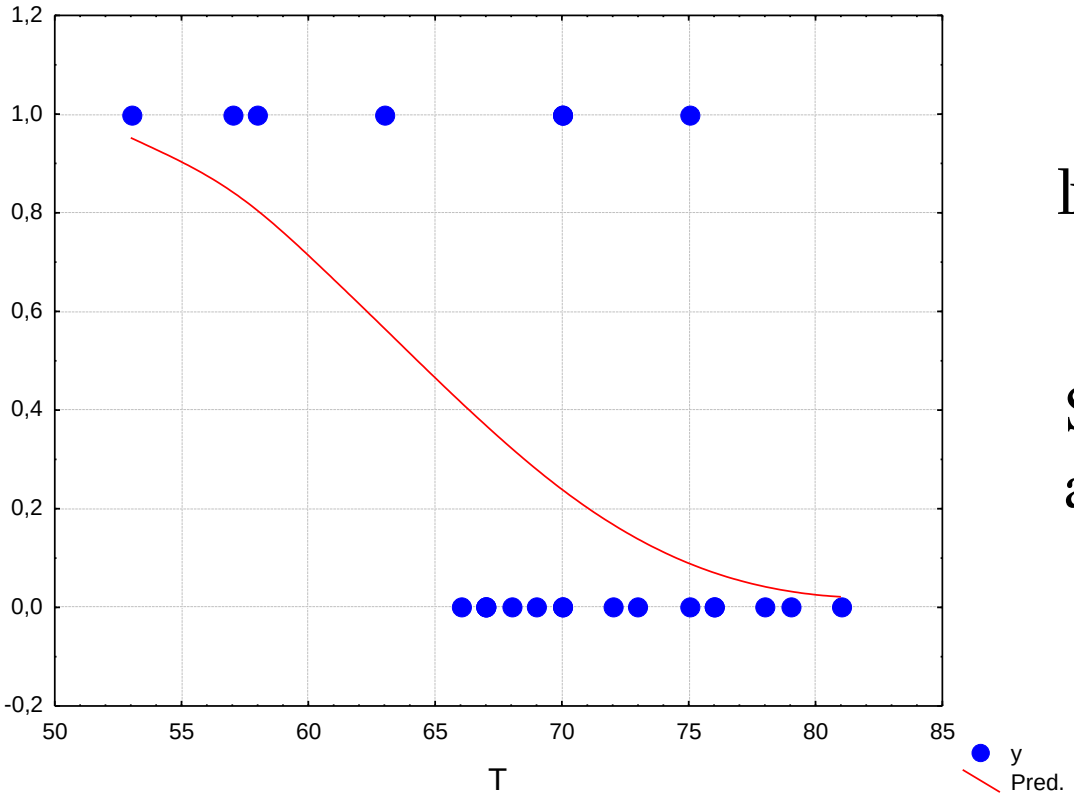
Dichotom (bináris) függő változó, folytonos független változó

1. példa

A. Agresti: Categorical data analysis, J. Wiley, 2002, p. 199

A Challenger űrrepülőgép katasztrófája (1986) után megvizsgálták, hogy a korábbi 23 repülés során mely esetekben károsodott a kritikusnak bizonyult O-gyűrű (1, ha igen, 0, ha nem), és ekkor milyen volt a külső hőmérséklet (^{0}F).


Scatterplot of y against T
challenger 10v*23c
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
y
50
55
60
65
70
75
80
85
T



$$\ln \frac{\pi}{1-\pi} = 15.043 - 0.232x$$

Szignifikáns
a hőmérséklet hatása?

$H_0: \beta = 0$

Effect	y - Parameter estimates (challenger) Distribution : BINOMIAL, Link function: LOGIT Modeled probability that y = 1 Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	15.0429(	7.378391	4.15661(	0.041472
T		2	-0.2321(	0.10823(	4.60115(	0.03195(
Scale			1.0000(	0.00000(		

Számítsuk ki a károsodás valószínűségét a baleset környezeti hőmérsékletére (31^0F)!

$$\ln\frac{\pi}{1-\pi} = 15.043 - 0.232x = 15.043 - 0.232 \cdot 31 = 7.851$$

$$\pi = \frac{\exp(\text{logit})}{1+\exp(\text{logit})} = \frac{\exp(7.851)}{1+\exp(7.851)} = 0.9996$$

Effect	y - Parameter estimates (challenger) Distribution : BINOMIAL, Link function: LOGIT Modeled probability that y = 1					
	Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	15.0429(	7.378391	4.15661(	0.041472
T		2	-0.2321(	0.10823(	4.601151	0.03195(
Scale			1.0000(	0.00000(		

Illeszkedés jósága – Pearson reziduum

Pearson-reziduum: általában $\dfrac{y_i - \hat{Y}_i}{\sqrt{Var(y_i)}}$ esetünkre: $\dfrac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$

Négyzetösszege (ha az illesztett egyenes adekvát) közelítőleg χ^2-eloszlást követ, n-paraméterek száma (egyenes esetén n-2) szabadsági fokkal.

A próbastatisztika: $\chi_0^2 = \sum_{i=1}^{n} \dfrac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$

	y - Statistics of goodness of fit (challe Distribution : BINOMIAL, Link function: Modeled probability that y = 1		
Stat.	Df	Stat.	Stat/Df
Deviance	21	20.3152	0.96739(
Scaled Deviance	21	20.3152	0.96739(
Pearson Chi²	21	23.1691	1.10329(
Scaled P. Chi²	21	23.1691	1.10329(

Illeszkedés jósága - deviancia

$$D = 2\ln\frac{L_{\max}}{L_{\text{modell}}} = 2\left(\ln L_{\max} - \ln L_{\text{modell}}\right)$$

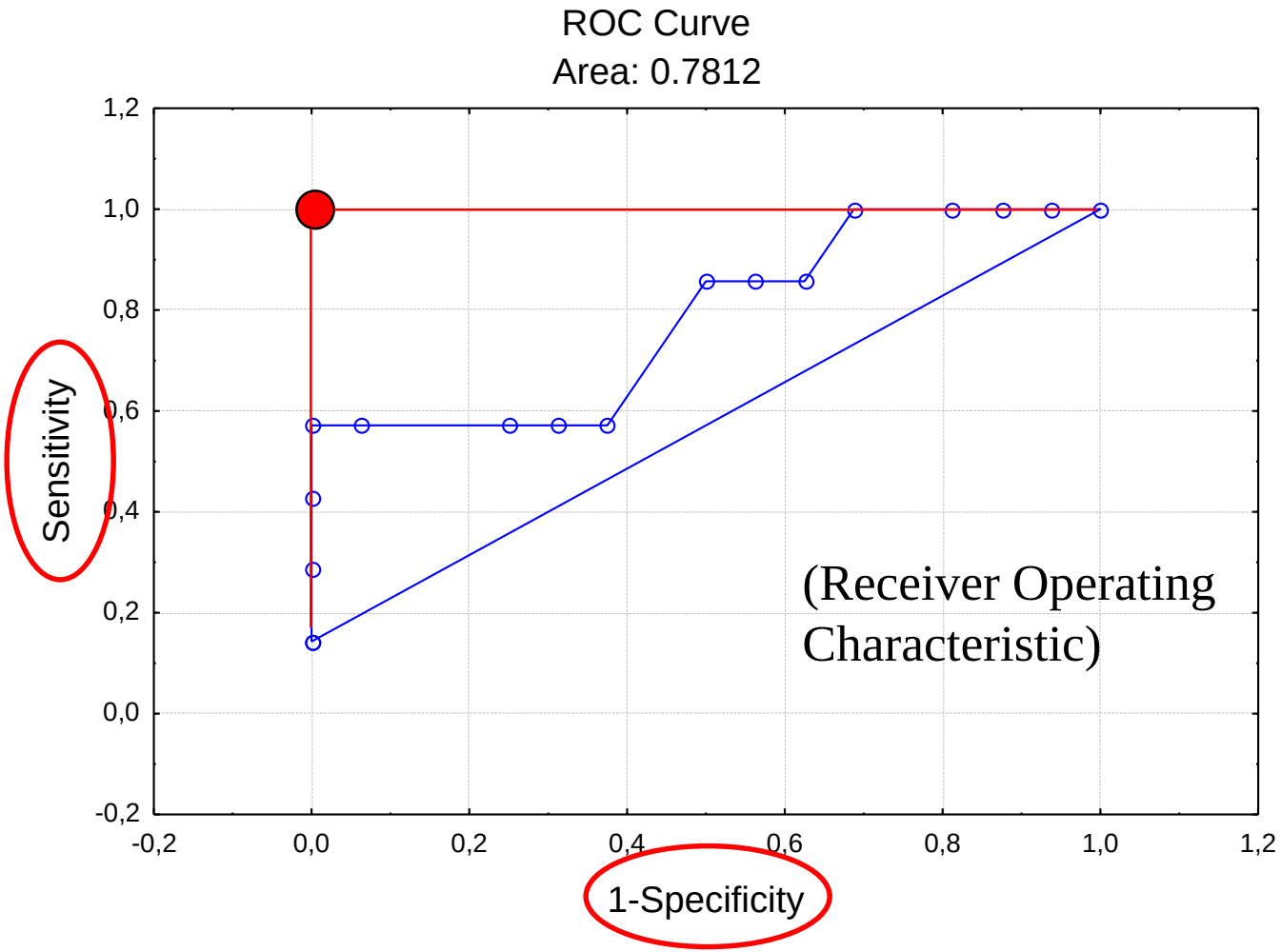
$$\ln L_{\max} = \sum_i \left[y_i \ln y_i + (1-y_i)\ln(1-y_i)\right]$$

$$\ln L_{\text{modell}} = \sum_i \left[y_i(a+bx_i) - \ln\left(1+e^{a+bx_o}\right)\right]$$

A deviancia is közelítőleg χ^2-eloszlást követ, egyenes esetén n-2 szabadsági fokkal

Stat.	y - Statistics of goodness of fit (challe Distribution : BINOMIAL, Link function: Modeled probability that y = 1		
	Df	Stat.	Stat/Df
Deviance	21	20.3152	0.96739(
Scaled Deviance	21	20.3152	0.96739(
Pearson Chi²	21	23.1691	1.10329(
Scaled P. Chi²	21	23.1691	1.10329(

Szenzitivitás: annak valószínűsége, hogy a modell az esemény bekövetkezését jósolja és tényleg bekövetkezik



Specificitás: annak valószínűsége, hogy a modell az esemény be nem következését jósolja és valóban nem következik be

Kell-e másodfokú függvény?

Effect	y - Likelihood Type 1 Test (challenger) Distribution : BINOMIAL Link function: LOGIT Degr. of Freedom	Log-Likelihd	Chi-Square	p
Intercept	1	-14.133(		
T	1	-10.157(	7.95196(	0.00480⁴
T^2	1	-9.694⁴	0.92649(	0.335777

Döntés?

2. példa

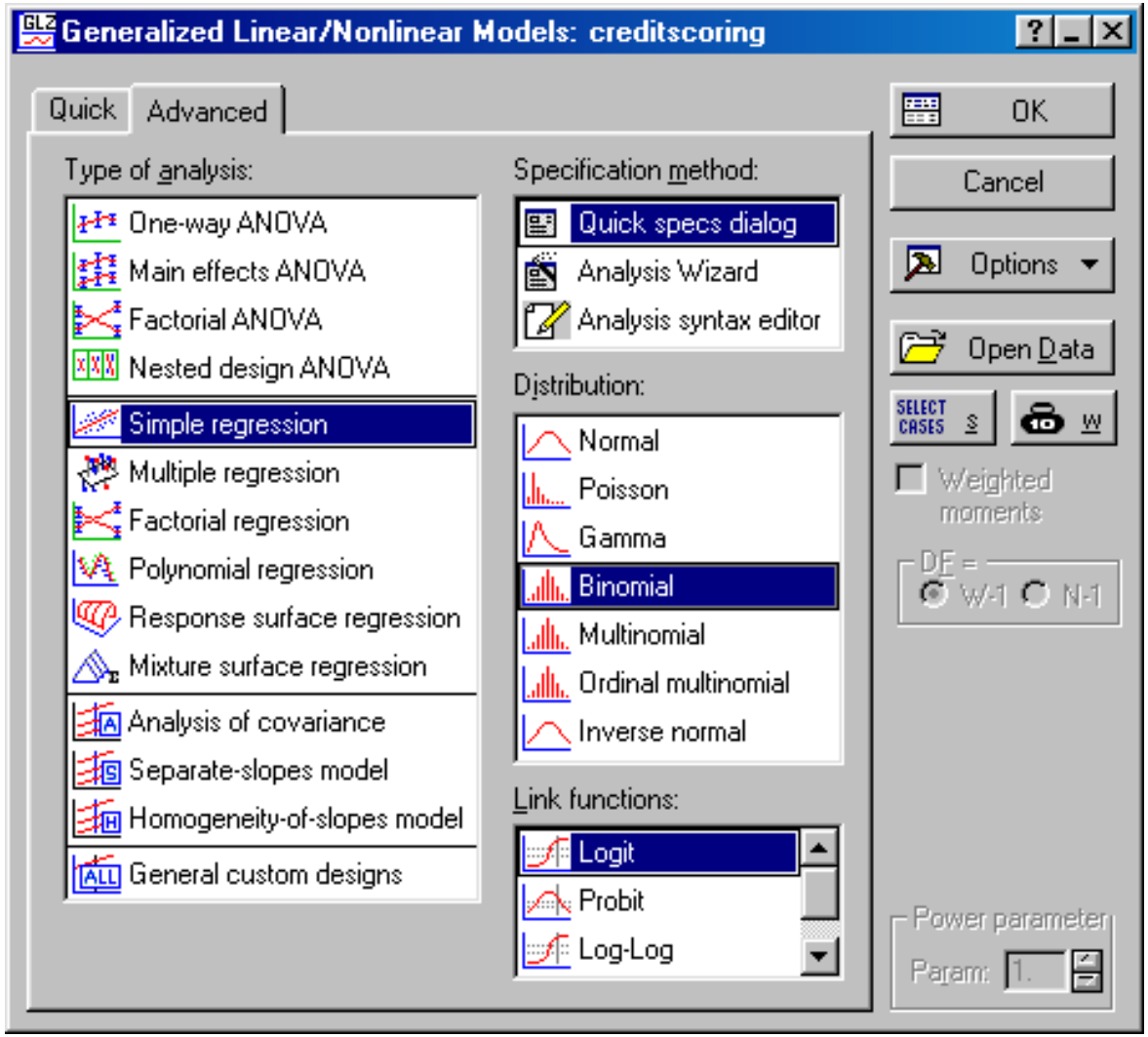
StatSoft példa,
source:http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html

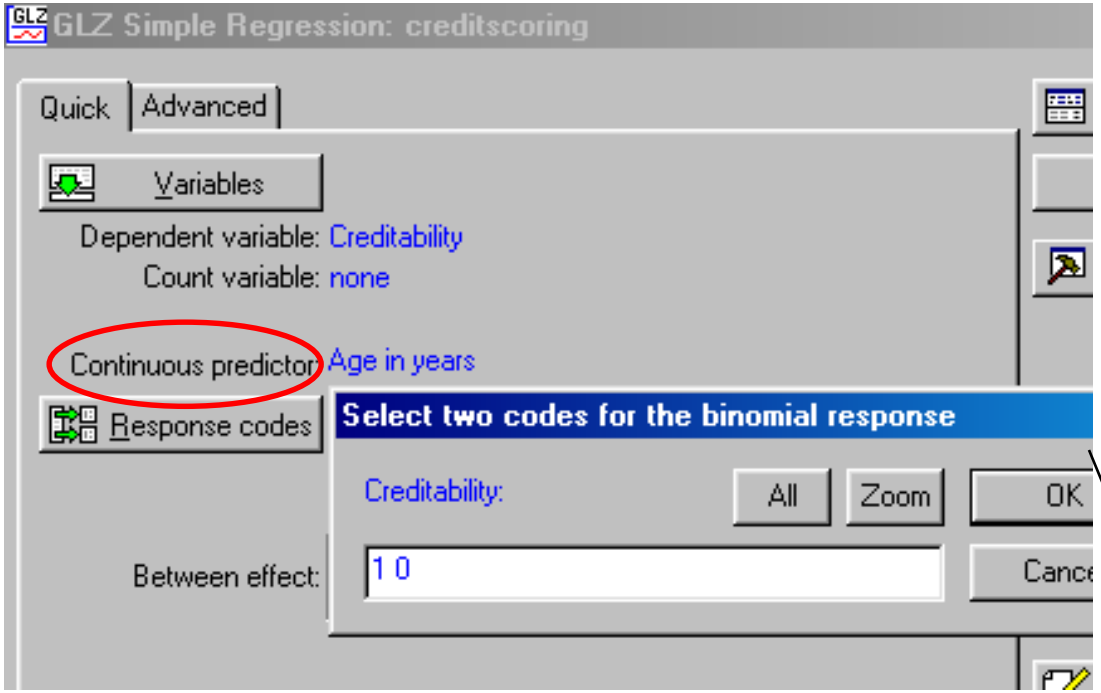
Credit scoring: a törlesztési hajlandóság függése az életkortól

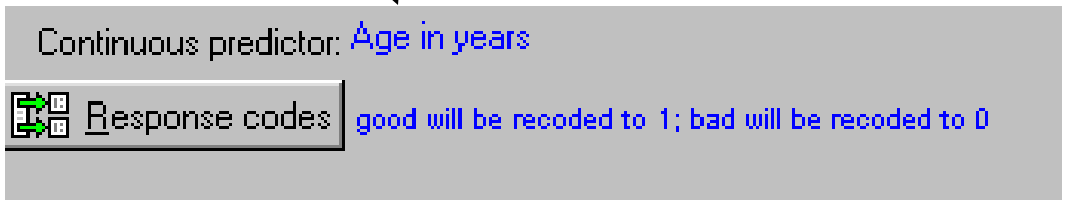
	1 Creditability	2 Age in years
1	bad	24
2	good	48
3	bad	26
4	good	44
5	good	25
6	good	39
7	bad	31

az adatfile részlete

Statistics >Advanced Linear/Nonlinear Models >
> **Generalized** Linear/Nonlinear Models







Estimates

Effect	Creditability - Parameter estimates (creditscoring) Distribution : BINOMIAL Link function: LOGIT Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	0.198456	0.233333	0.723402	0.395030
Age in years		2	0.018512	0.006449	8.239067	0.004100
Scale			1.000000	0.000000		

$$\text{logit} = \ln \frac{\pi}{1-\pi} = \alpha + \beta x$$

Dichotom függő változó, névleges skálán mért független változó

A függő és a független változó is csak kétféle értéket vehet föl

pl. x kétféle értéket vehet föl (van-e telefonja vagy sem),
azaz x=0 vagy x=1

$$\ln \frac{\pi}{1-\pi} = \alpha + \beta x$$

A β paraméter jelentése: ennyivel változik az esély logaritmusa,
ha x 0-ról 1-re változik

3. példa

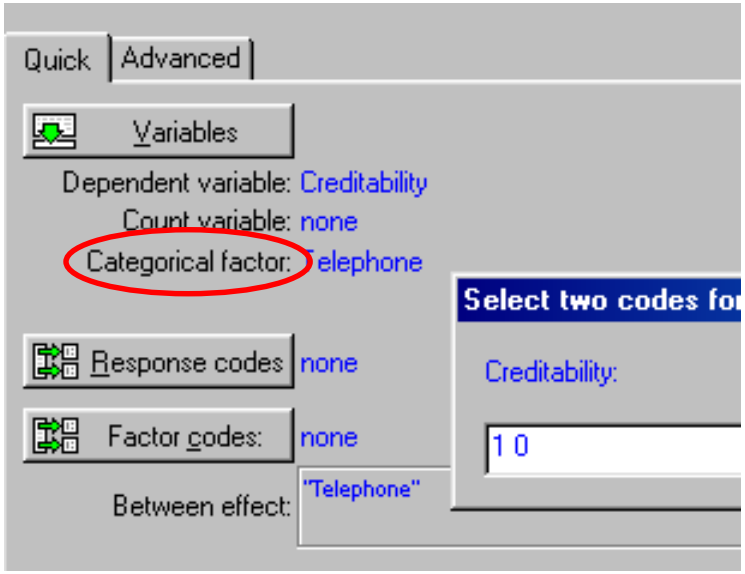
StatSoft példája

source:http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html

Credit scoring: a törlesztési hajlandóság függ-e attól, hogy van-e telefonja

	1 Creditability	2 Telephone
1	bad	yes
2	good	yes
3	bad	yes
4	good	yes
5	good	yes
6	good	no
7	bad	no

az adatfile részlete



Effect	Creditability - Parameter estimates (creditscoring) Distribution : BINOMIAL Link function: LOGIT Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	0.782607	0.088276	78.59693	0.000000
Telephone	yes	2	0.163329	0.141699	1.32860	0.249055
Telephone	no	3	0.000000			
Scale			1.000000	0.000000		

$$\text{logit} = \ln \frac{\pi}{1-\pi} = \alpha + \beta x$$

$\hat{Y} = 0.7826$ ha nincs telefonja

β-val változik az ln esélyhányados, ha x $0 \rightarrow 1$ –re változik (nincs $\rightarrow$ van)

$\hat{Y} = 0.7826 + 0.1633 = 0.9459$ ha van

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\hat{\pi} = \frac{e^{\text{logit}}}{1+e^{\text{logit}}} = \frac{e^{0.7826}}{1+e^{0.7826}} = 0.686$$ ha nincs telefonja

$$\hat{\pi} = \frac{e^{0.9459}}{1+e^{0.9459}} = 0.720$$ ha van

Dichotom függő változó, több független változó

$$Y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_r x_{ri}$$

Mindegyik *x* lehet folytonos vagy diszkrét, vegyesen is lehet

4.1. példa

A tumor-átmérő (x) és az 5 éves túlélés kimenetele (y) közötti összefüggést vizsgálták 181 beteg adataiból.

Effect	tul5evm - Parameter estimates (DRPETE3f) Distribution : BINOMIAL Link function: LOGIT Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	1.872815	0.321752	33.88027	0.000000
TUATM		2	-0.019812	0.008499	5.43424	0.019746
Scale			1.000000	0.000000		

$$\ln\frac{\pi}{1-\pi} = 1.873 - 0.0198x$$

azaz ha a tumor átmérője egy egységgel nő, az 5 éves túlélés esélyének logaritmusa kb. 0.02-dal csökken

Konfidencia-intervallum a paraméterekre:

Effect	tul5evm - Confidence Intervals of Estimates (DRPETE3 Distribution : BINOMIAL Link function: LOGIT Level of Effect	Column	Lower CL 95. %	Upper CL 95. %
Intercept		1	1.242193	2.503438
TUATM		2	-0.036469	-0.003155

$$\ln\frac{\pi}{1-\pi} = 1.873 - 0.0198x$$

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Kérdések:

Mennyi az 5 éves túlélés valószínűsége, ha a tumorátmérő 0 mm?

Mennyi az 5 éves túlélés valószínűsége, ha a tumorátmérő 38 mm?

Mekkora tumorátmérőnél lesz az 5 éves túlélés valószínűsége 40%?

Hányszor nagyobb az 5 éves túlélés valószínűsége, ha a tumorátmérő 40 mm?

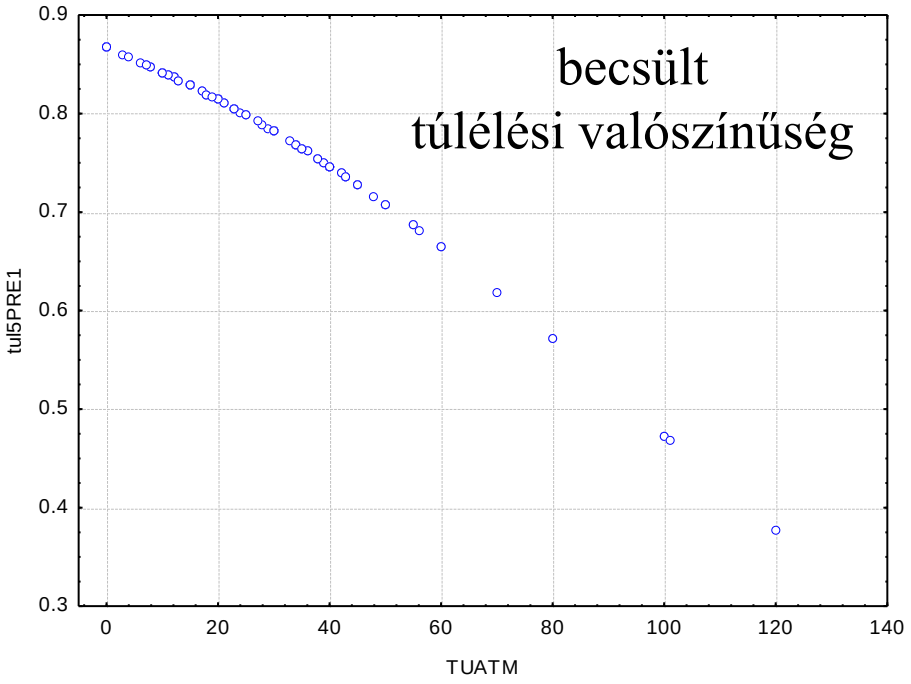
Modell „jósságának” vizsgálata:

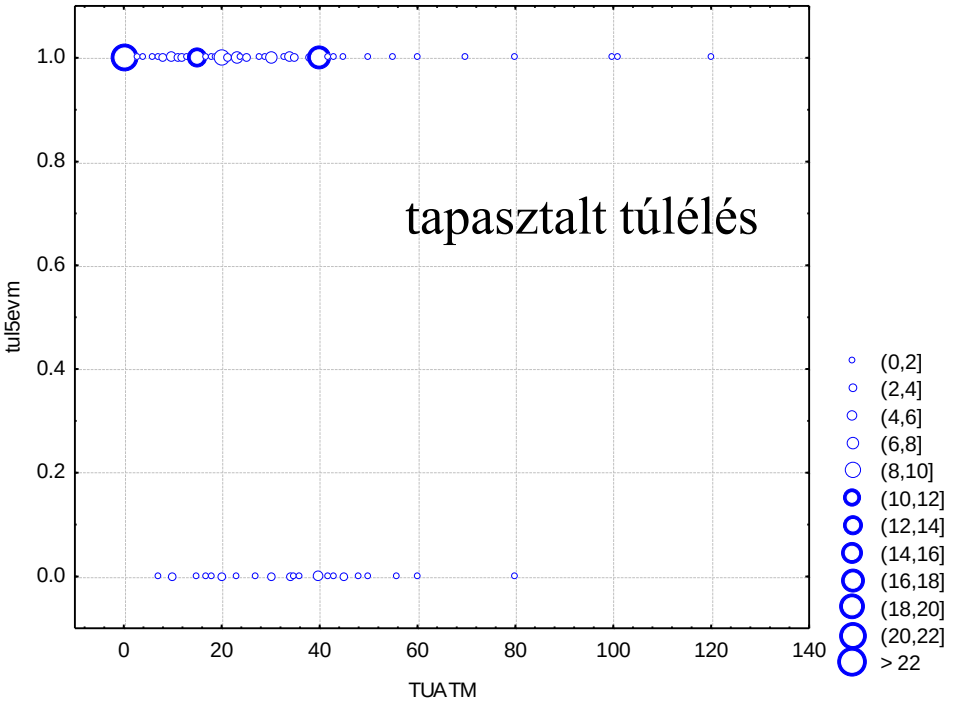
A tapasztalati és becsült esélyhányados:

	Classification of cases (DRPETE3f Odds ratio: 0.000000 Log odds ratio: infinity		
Observed	Predicted 1	Predicted 0	Percent correct
1	138	4	97.1831(
0	39	0	0.0000(

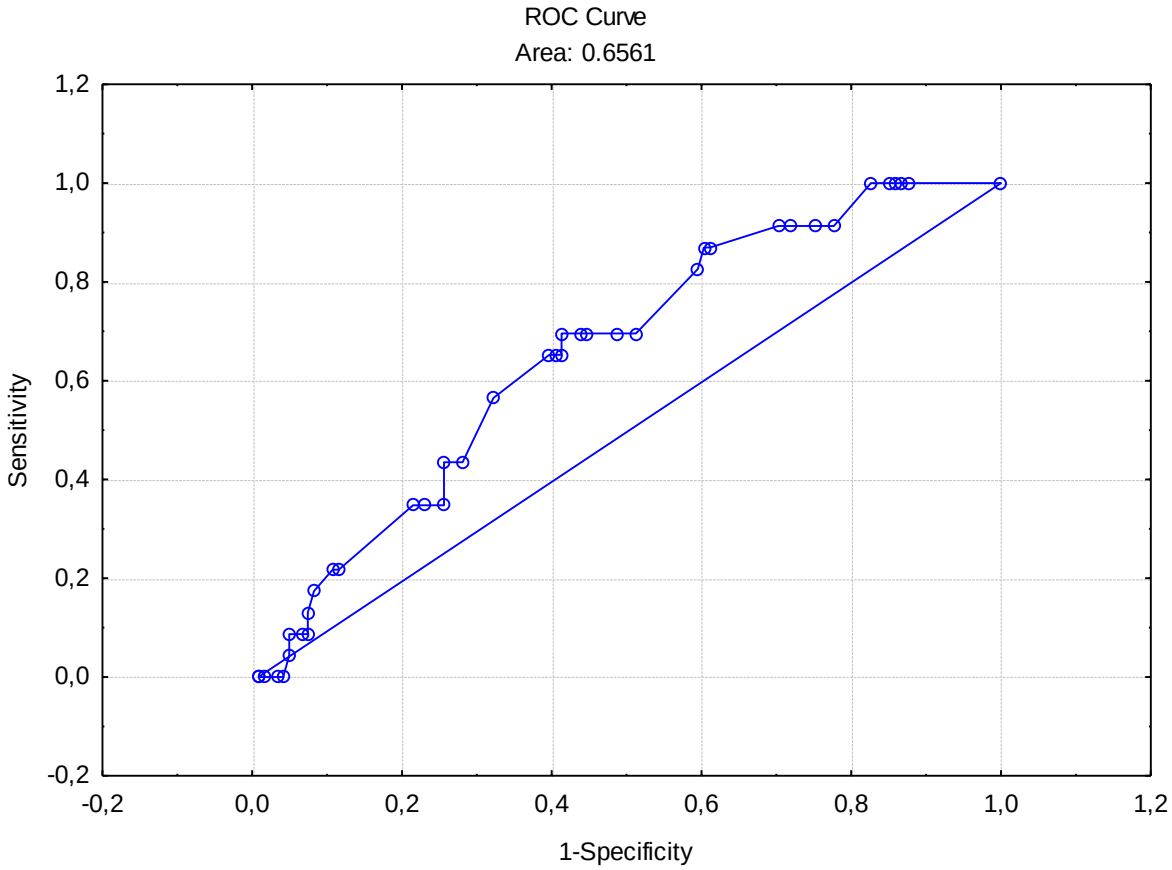
$\hat{\pi} > 0.5 \rightarrow$ túléltnek becsüljük

főátlóbeli elemeknek kell nagynak lennie,
azaz ennek a modellnek a magyarázó ereje csekély





Modell „jósáságának” vizsgálata:



4.2. példa

Vizsgáljuk most azt a modellt, mely a tumor-átmérő (x_1), az ér-nyirokér-áttét léte (x_2) és az 5 éves túlélés kimenetele (y) közötti összefüggést írja le (x_2=0, ha nem volt ilyen áttét, x_2=1, ha volt).

tul5evm - Parameter estimates (DRPETE3f)
Distribution : BINOMIAL
Link function: LOGIT

Effect	Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	1.596414	0.338217	22.27928	0.000002
TUATM		2	-0.018767	0.008594	4.76834	0.028988
ernyirokm	0	3	0.537129	0.201816	7.08350	0.007780
Scale			1.000000	0.000000		

Classification of cases (DRPETE3f.sta
Odds ratio: 1.842105
Log odds ratio: 0.610909
Include condition: exit1<>6

Observed	Predicted 0	Predicted 1	Percent correct
0	1	38	2.56410
1	2	140	98.59155

4.3. példa

Vegyük bele az előbbi modellbe a tumor-átmérő (x_1), az érnyirokér-áttét léte (x_2) közti kölcsönhatást is!

tul5evm - Parameter estimates (DRPETE3f.sta)
Distribution : BINOMIAL, Link function: LOGIT
Modeled probability that tul5evm = 0
Include condition: exit1<>6

Effect	Level of Effect	Column	Estimate	Standard Error	Wald Stat.	Lower CL 95. %	Upper CL 95. %	p
Intercept		1	-1.8380(	0.433553	17.97247	-2.68775	-0.98825‹	0.000022
ernyirokm	0	2	-0.1621(	0.433553	0.13979	-1.01185	0.687651	0.708492
TUATM		3	0.02607	0.011711	4.9549(	0.00312	0.049022	0.026017
ernyirokm*TUATM	0	4	-0.0116[	0.011711	0.98939	-0.0346(	0.011305	0.319893
Scale			1.0000(	0.00000(		1.0000(	1.00000(	

Classification of cases (DRPETE3f.sta
Odds ratio: 12.727273
Log odds ratio: 2.543747
Include condition: exit1<>6

Observed	Predicted 0	Predicted 1	Percent correct
0	6	33	15.38462
1	2	140	98.59155