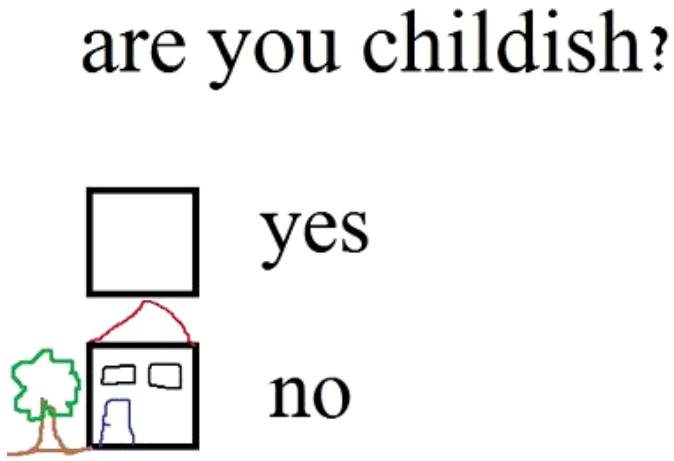
Módszertani készségfejlesztés, többváltozós statisztikai eljárások (BMNPS16100M)

Készítette: Soltész-Várhelyi Klára

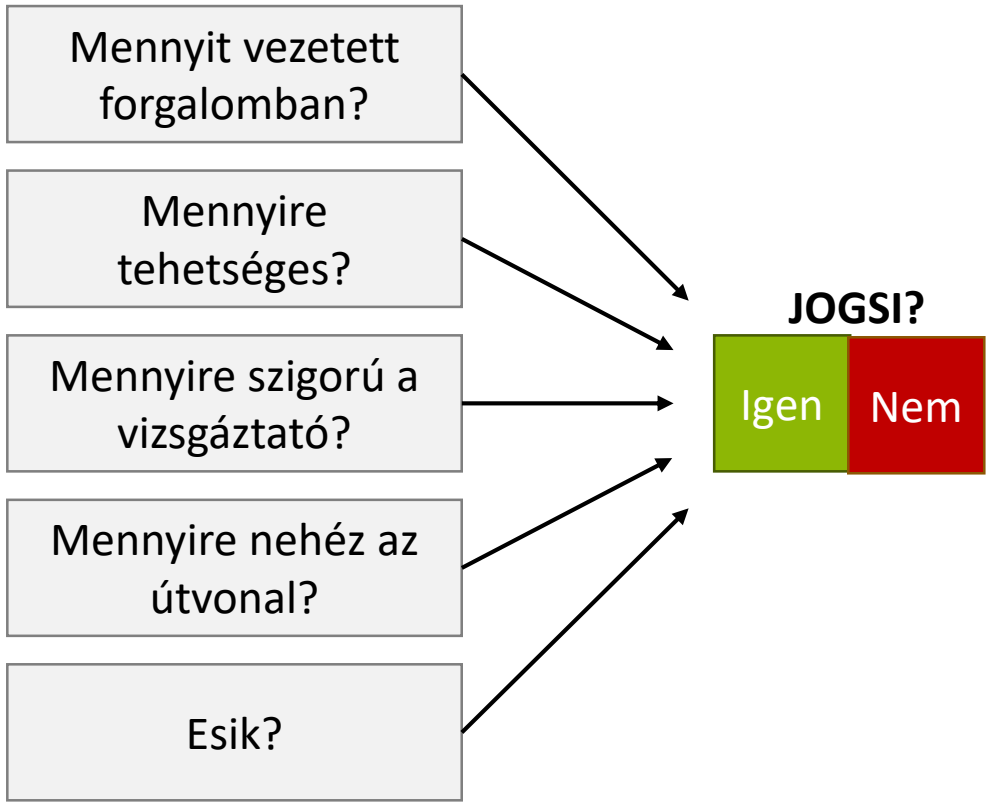
Logisztikus regresszió

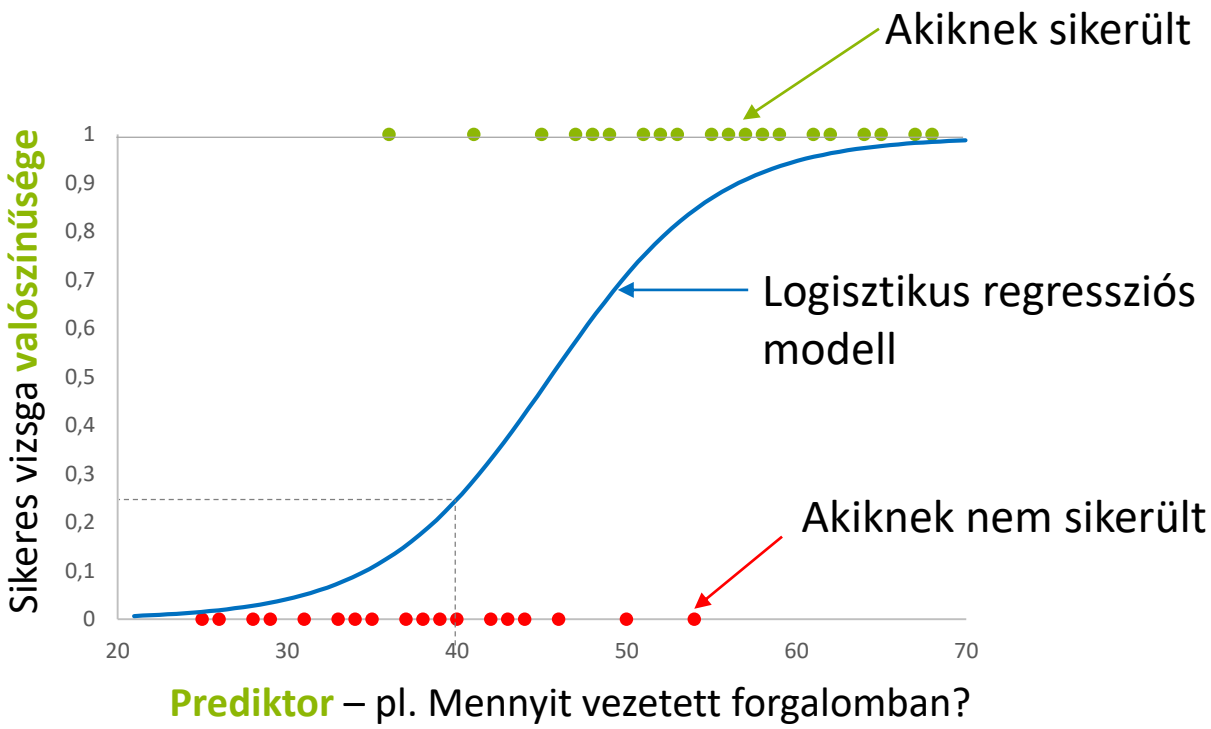
Bináris logisztikus regresszió elmélete



Példa, mire használjuk

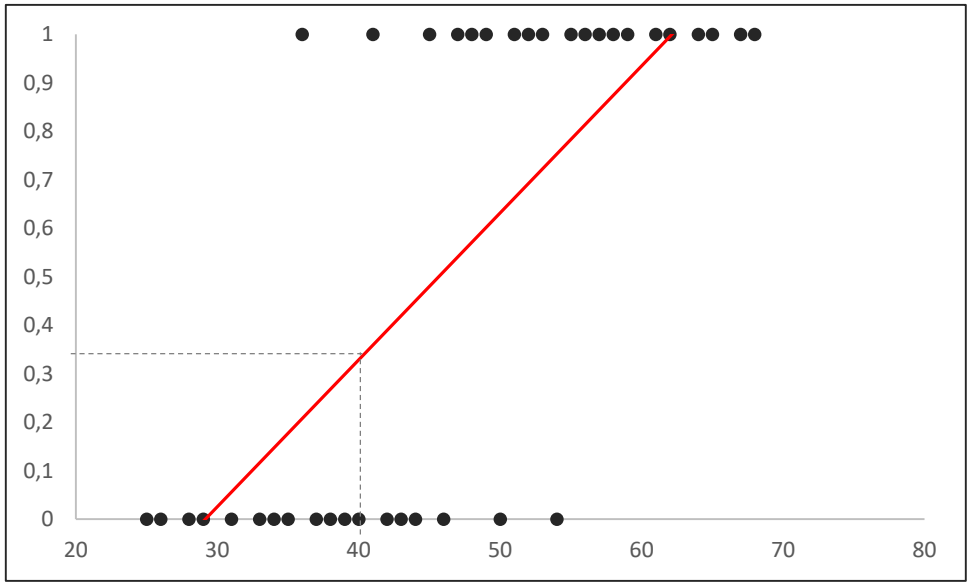
- Lineáris regresszióhoz hasonlóan prediktor változók segítségével próbáljuk bejósolni egy kimeneti változó értékét, de a kimeneti változó most dichotóm (bináris)
- Milyen prediktorokból, milyen mértékben jósolható be, hogy…
  - …a műtét sikeres lesz-e? …a hitelt kérő fizeti-e majd a kölcsönt? …sikerül-e átmenni a vizsgán? …meglesz-e a jogsi?





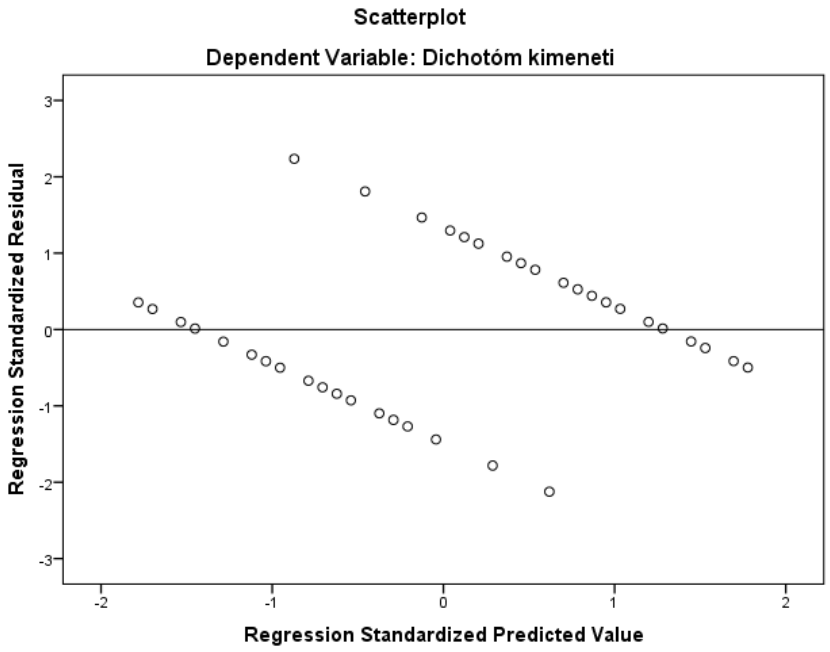
Lineáris regresszió dichotóm adatokra?

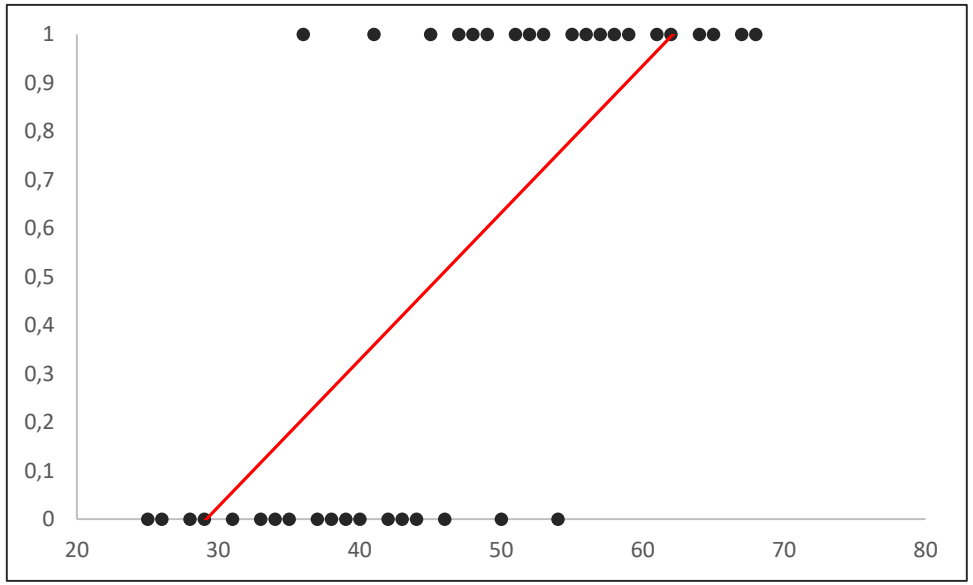
- Miért nem működik ez lineáris regresszióval?
- A regressziós vonal az 1-es kimenet (azaz egy esemény megtörténtének) valószínűségét jelentené
  - a példában a 40-as prediktor változó érték esetén annak valószínűsége, hogy a kimeneti változó értéke 1 lesz, kb. 33%
  - Ezzel a modellel azonban több probléma is van



Probléma a lineáris regresszióval

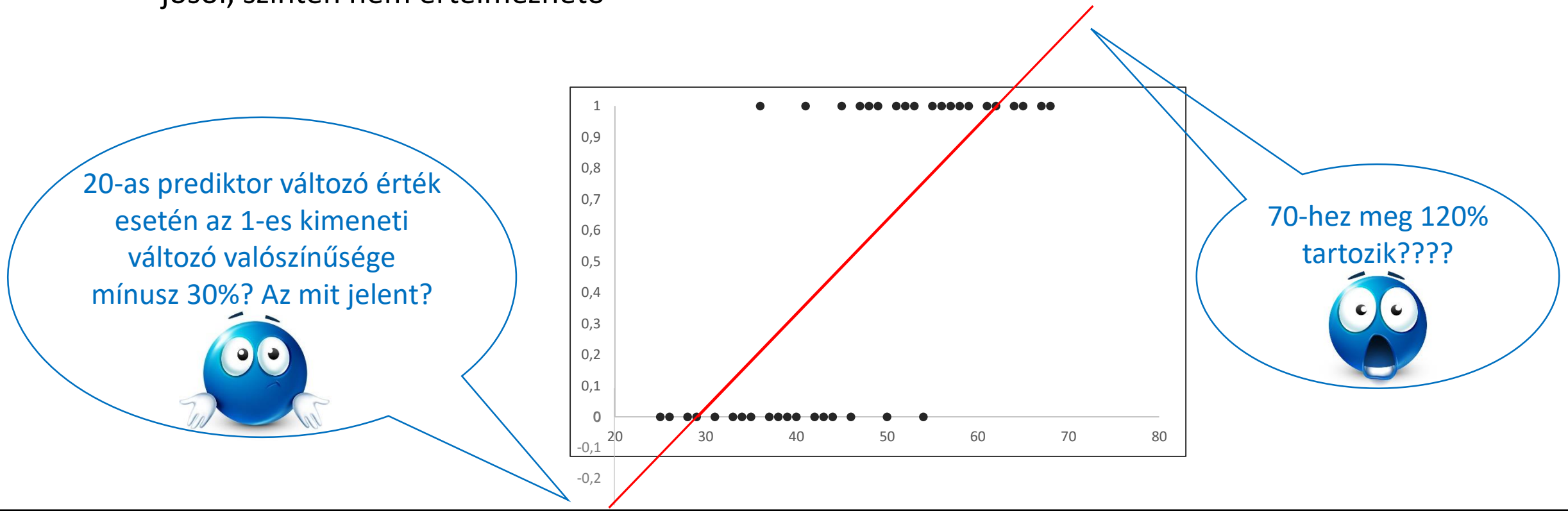
- A bináris kimeneti változó a lineáris regresszió számos feltételét sérti:
  - A kimeneti változó nem skála típusú
  - Nem teljesül a hibák normalitása
  - Nem teljesül a szóráshomogenitás





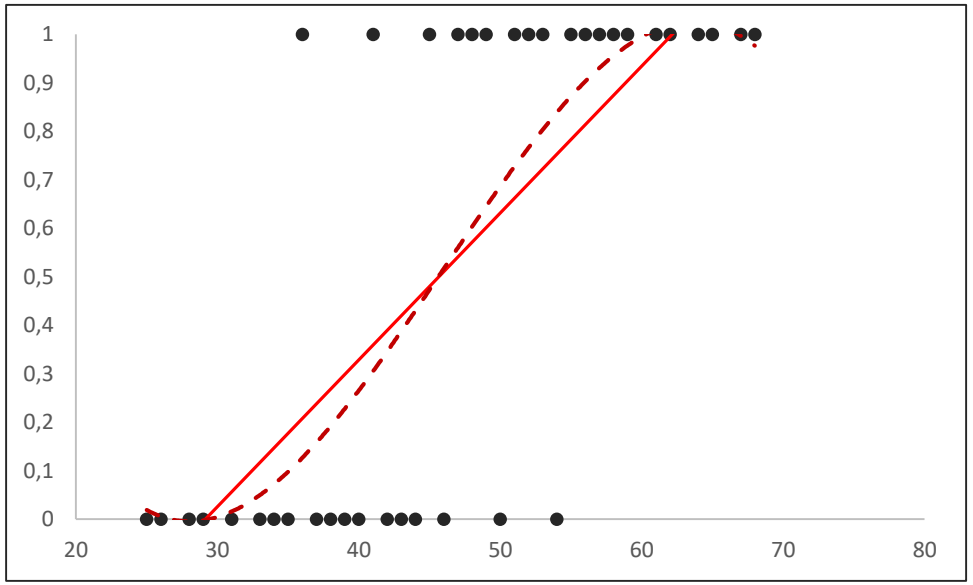
Probléma a lineáris regresszióval

- További probléma, hogy a modell nem extrapolálható (alkalmazható a minta értékeinél szélsőségesebbre)
  - Ha az egyenest tovább húznánk, pl. 30 alatti prediktor változó értékhez szeretnénk bejósolni a kimeneti változó valószínűségét, akkor a modell negatív valószínűséget jósol, ami nem értelmezhető
  - Ha 60 fölötti prediktor változó értékhez szeretnénk jósolni, akkor pedig 100% feletti valószínűséget jósol, szintén nem értelmezhető



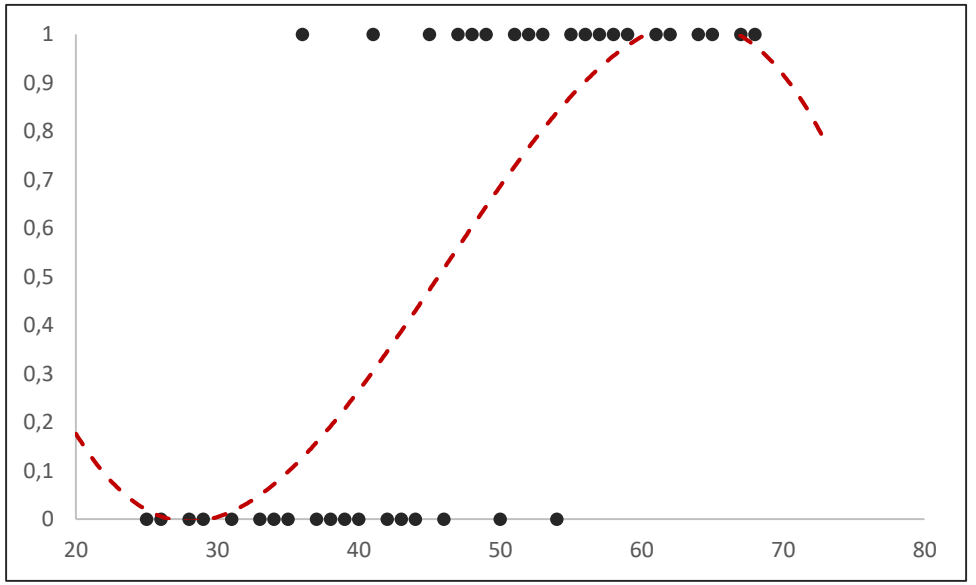
Probléma a lineáris regresszióval

- További probléma, hogy az egyenes nem írja jól le a tényleges valószínűséget
  - A tényleges valószínűség sokkal inkább egy S alakú görbe mentén halad, alacsony prediktor értékek esetén az 1-es valószínűsége sokáig alacsony, majd egy viszonylag gyors váltás után a magas prediktor értékekhez az 1-es kimeneti érték valószínűsége magas



Probléma a lineáris regresszióval

- Nem működik az extrapoláció a harmadrendű modellel sem
  - Ha egy harmadrendű (S alakú) modellt használnánk, akkor pedig a modell a nagyon alacsony és nagyon magas értékek esetén „visszafordulna”, és például a nagyon magas prediktor változó értékhez megint csak alacsonyabb valószínűséget predikálna.



Miért jobb az esély, mint a valószínűség?

- A probléma megoldható egy kis trükkel, az esélyek bevezetésével
- **Valószínűség** (probability)

$$p = \frac{\text{kívánt kimenet}}{\text{összes lehetséges kimenet}} \qquad q = \frac{\text{nem kívánt kimenet}}{\text{összes lehetséges kimenet}}$$

  - Pénzfeldobásnál, ha fejre fogadtam, akkor a kívánt kimenet a fej, az összes kimenet a fej és írás, tehát annak a valószínűsége, hogy nyerek, azaz hogy fej lesz $p = \frac{1}{2} = 0.5$ azaz 50%
  - Dobókockával 1/6 = 16% valószínűsége van annak, hogy hatost dobsz elsőre

- **Esély** (odds)

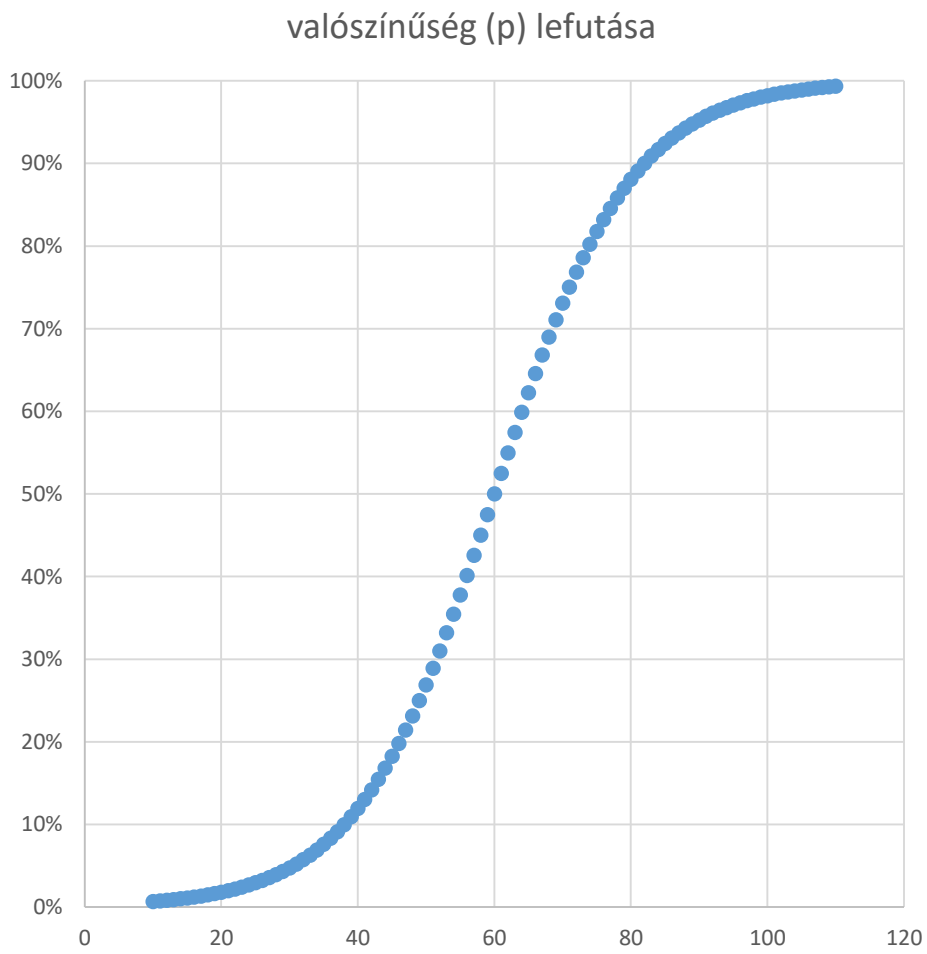
$$\text{odds} = \frac{\text{kívánt kimenet}}{\text{nem kívánt kimenet}} = \frac{p}{q}$$

  - pénzfeldobásnál 1 az 1-hez a fej esélye, azaz p = 50%, és q=50%, tehát az esély 1
  - 1 az 5-höz az esélye annak, hogy hatost dobsz, tehát az esély: odds = $\frac{p}{q} = \frac{1/6}{5/6} = \frac{1}{5} = 0.2$

Miért jobb az esély, mint a valószínűség?

- **p(Y)** annak valószínűsége, hogy Y beteljesül (értéke 1 lesz)

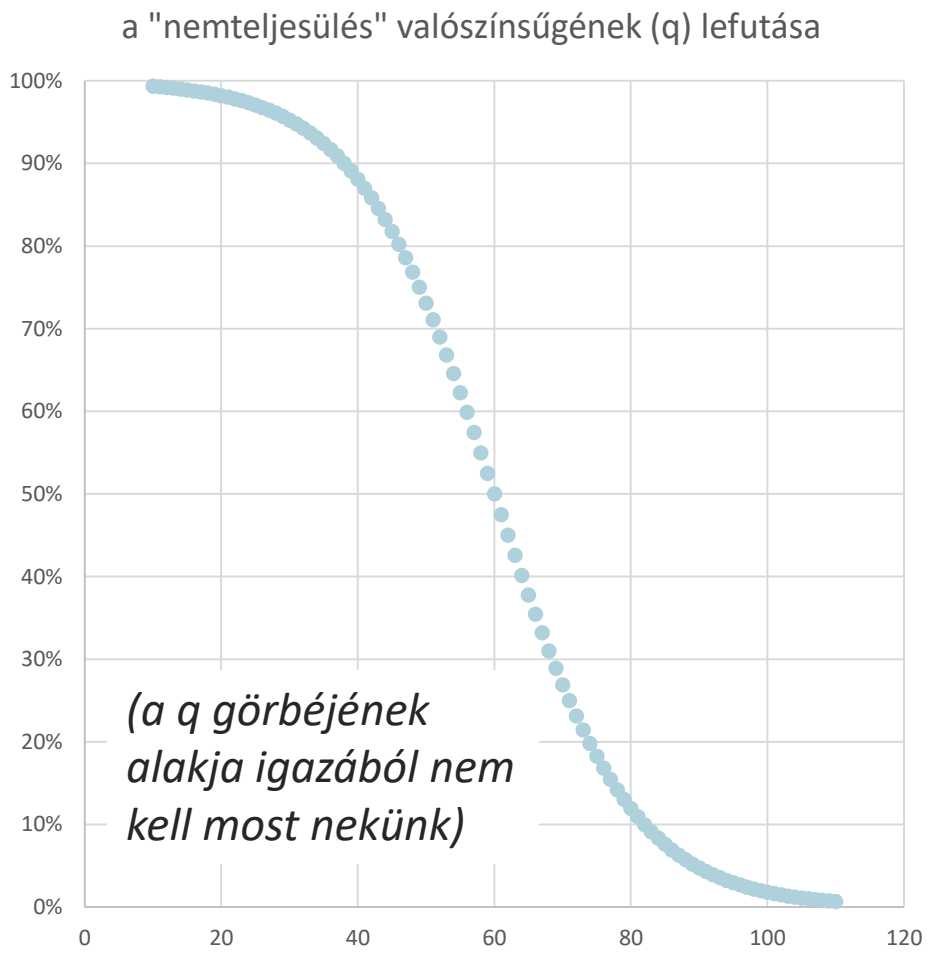
x	p
11	0,0074
12	0,0082
13	0,0090
20	0,0180
30	0,0474
40	0,1192
50	0,2689
60	0,5000
70	0,7311
80	0,8808
90	0,9526
100	0,9820
107	0,9910
108	0,9918
109	0,9933



Miért jobb az esély, mint a valószínűség?

- **q(Y)** annak valószínűsége, hogy Y nem teljesül be, Y értéke 0 lesz. Számítása: q = 1 - p

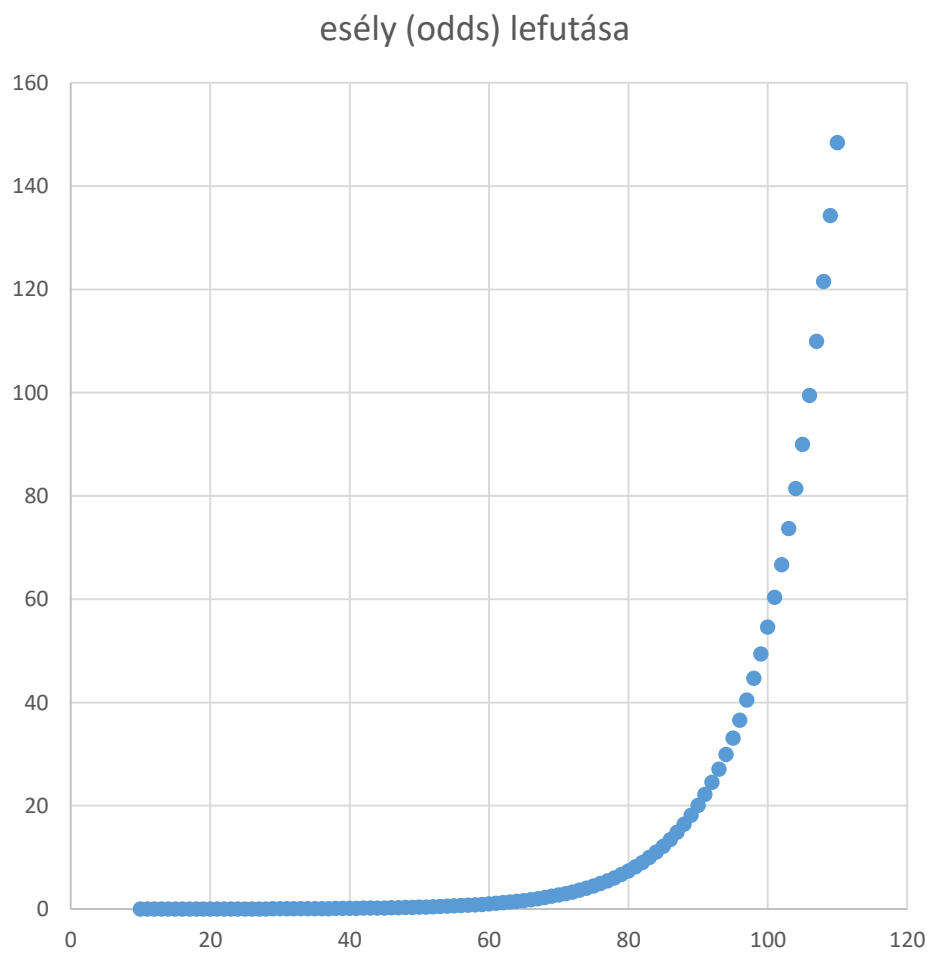
x	p	q
11	0,0074	0,9926
12	0,0082	0,9918
13	0,0090	0,9910
20	0,0180	0,9820
30	0,0474	0,9526
40	0,1192	0,8808
50	0,2689	0,7311
60	0,5000	0,5000
70	0,7311	0,2689
80	0,8808	0,1192
90	0,9526	0,0474
100	0,9820	0,0180
107	0,9910	0,0090
108	0,9918	0,0082
109	0,9933	0,0067



Miért jobb az esély, mint a valószínűség?

- **odds(Y)** annak esélye, hogy Y beteljesül. Számítása: odds = p/q (próbáld levezetni, miért ez a képlet!). Előnye: az odds értéke akármilyen magas lehet, nincs felső határa, mint a valószínűségnek.

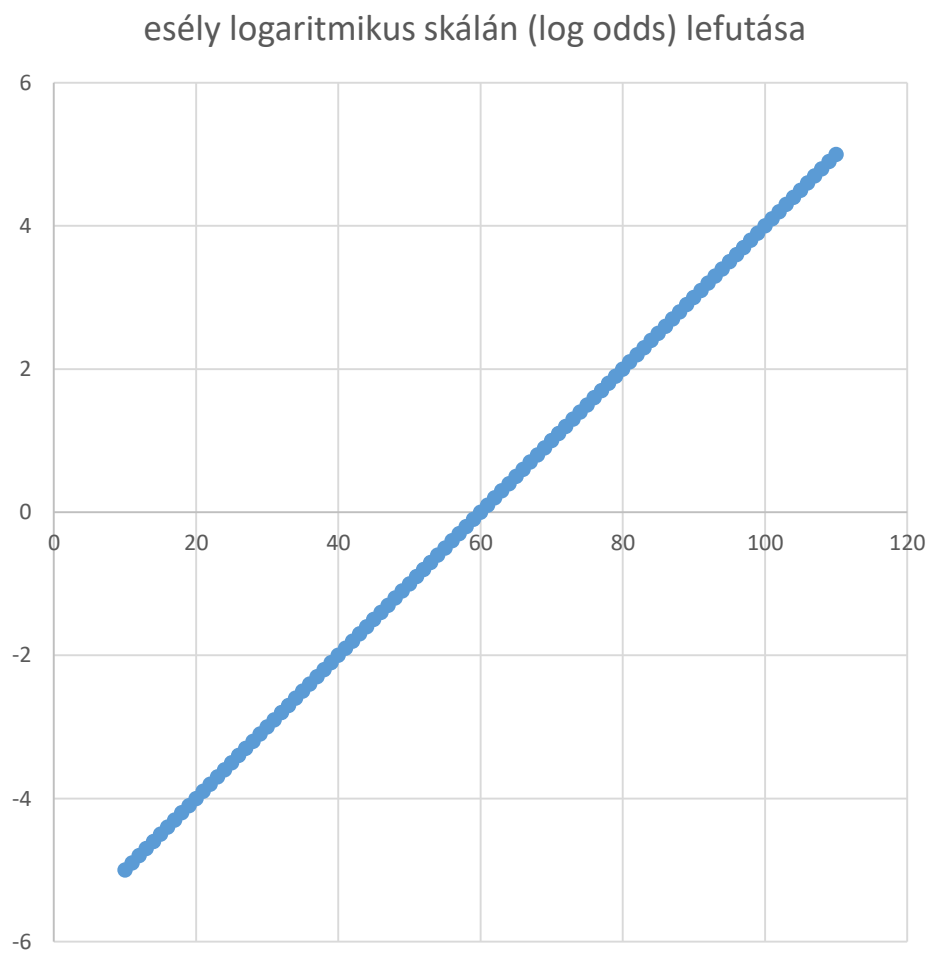
x	p	q	odds
11	0,0074	0,9926	0,0074
12	0,0082	0,9918	0,0082
13	0,0090	0,9910	0,0091
20	0,0180	0,9820	0,0183
30	0,0474	0,9526	0,0498
40	0,1192	0,8808	0,1353
50	0,2689	0,7311	0,3679
60	0,5000	0,5000	1,0000
70	0,7311	0,2689	2,7183
80	0,8808	0,1192	7,3891
90	0,9526	0,0474	20,0855
100	0,9820	0,0180	54,5982
107	0,9910	0,0090	109,9472
108	0,9918	0,0082	121,5104
109	0,9933	0,0067	148,4132



Miért jobb az esély, mint a valószínűség?

- **Ln(odds)** – az esély természetes logaritmusa. Előnyei: (1) Lehet akármilyen magas (2) vagy lehet negatív szám is, nincs felső és alsó határa, mint a valószínűségnek. (3) A legtöbbször megközelítőleg jól leírható egy egyenessel.

x	p	q	odds	ln(odds)
11	0,0074	0,9926	0,0074	-4,9
12	0,0082	0,9918	0,0082	-4,8
13	0,0090	0,9910	0,0091	-4,7
20	0,0180	0,9820	0,0183	-4
30	0,0474	0,9526	0,0498	-3
40	0,1192	0,8808	0,1353	-2
50	0,2689	0,7311	0,3679	-1
60	0,5000	0,5000	1,0000	0
70	0,7311	0,2689	2,7183	1
80	0,8808	0,1192	7,3891	2
90	0,9526	0,0474	20,0855	3
100	0,9820	0,0180	54,5982	4
107	0,9910	0,0090	109,9472	4,7
108	0,9918	0,0082	121,5104	4,8
109	0,9933	0,0067	148,4132	5



Maximum Likelihood

- A logaritmizált esély használata lehetővé teszi, hogy egy egyenest használjunk modellként a számolások során – ami matematikailag sokkal egyszerűbb. Most már csak (mint a lineáris regresszióban is) meg kell találnunk azt az egyenest, ami legjobban leírja az összefüggést.

- **Maximum Likelihood**
  - A logisztikus regresszió a modell megtalálásához a Legkisebb négyzetek helyett Maximum likelihood módszert használ.
    - A Legkisebb négyzetek módszere a modell predikciójának hibáját próbálja minimalizálni
    - A Maximum likelihood módszer a találati arányt próbálja meg maximalizálni
    - A két megfogalmazás konceptuálisan tautologikus – ha a hiba csökken, a találat nő
    - Számolásban azonban máshogy működik a kettő

	Hagyományos regresszió	**Logisztikus regresszió**
Modell megtalálásának módja	Legkisebb négyzetek módszere	Maximum likelihood
Konceptuálisan	Azt a modellt keressük, ahol a legkisebb a tévedés	Azt a modellt keressük, ahol a legnagyobb a találati arány
Számolásban	Algoritmikus megoldáskeresés	Próba & hiba módszer

Logisztikus regresszió egyenlete

- Ha megvan a legjobb egyenes, már csak vissza kell számoljunk a logaritmizált esélyekből a valószínűségeket (ez egyszerű egyenletrendezés), és máris kialakítottuk a **logisztikus regresszióra jellemző** S alakú **modell**t, **amellyel becsülhető a kimeneti érték valószínűsége**

  - A hagyományos lineáris regresszió képlete:

$$Y_i = b0 + b1 * X_i$$

  - A logisztikus regresszió képlete (ahol a p az esemény bekövetkezésének valószínűsége):

$$p(Y_i) = \frac{1}{1+ e^{-(b0 + b1 * X_i)}}$$

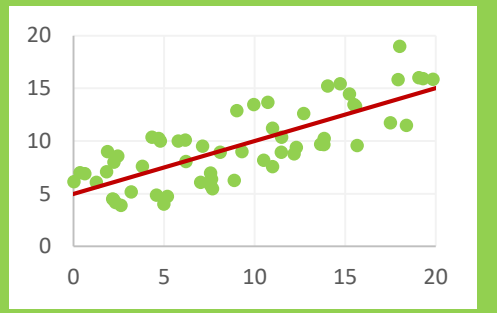
  - A **mért érték (Y_i)** csak 0 (nem következett be) és 1 (bekövetkezett) lehet
  - A **predikált érték a $p(Y_i)$**, azaz **az esemény bekövetkezésének valószínűsége**
    - 0% jelentése, hogy biztos nem következik be
    - 100% jelentése, hogy biztos bekövetkezik

Eddigiek összegzése

Lineáris regresszió

Folytonos kimeneti változó

Folytonos vagy dichotóm prediktorok



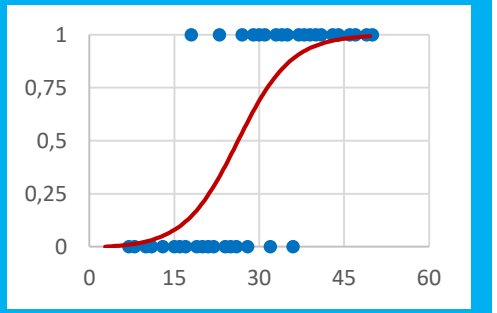
A modell:

$$Y_{pred,i} = b0 + b1 * X_i$$

Logisztikus regresszió

Dichotóm kimeneti változó

Folytonos, dichotóm vagy ordinális prediktorok



A modell kicsit bonyolultabb:

$$p(Y_i) = \frac{1}{1 + e^{-(b0 + b1 * X_i)}}$$

-2LL

- **Találat** számolása:
  - Hogy mennyire pontos a modell, azt itt is a predikált és mért értékek közötti együttjárás erősségéből tudjuk megállapítani (a tényleges számolása ennél kicsit bonyolultabb, a predikált és tényleges kimenetek valószínűségének összegzésén alapul)
  - Logisztikus regresszióban ezt **log-likelihoodnak (LL)** nevezzük.

- A hatásból a **tévedés** (eltérés a predikció és mért értékek között) könnyen számolható

  Eltérés = -2 * LL

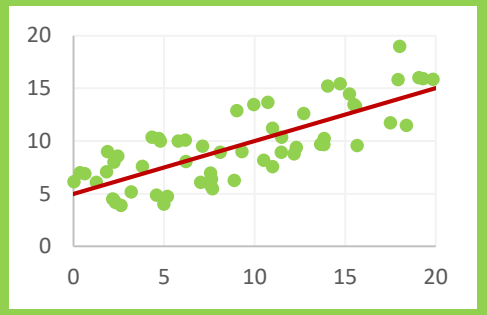
  - A tévedésnek nincs külön jelölése, általában **-2LL** néven hivatkozunk rá.
  - **Az eltérés** fontos tulajdonsága, hogy tudjuk, hogy **χ^2 (khi-négyzet) eloszlást követ**
  - Bár mindkettő működne, a hatás és tévedés közül általában a tévedéssel dolgozunk.

Eddigiek összegzése

Lineáris regresszió

Folytonos kimeneti változó

Folytonos vagy dichotóm prediktorok



A modell:

$$Y_{pred,i} = b0 + b1 * X_i$$

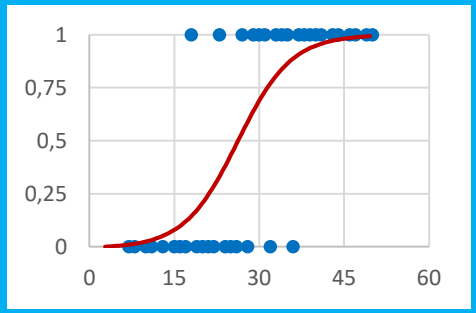
Modell pontossága?

SS_R, azaz a **legkisebb négyzetek módszer**ével számolt eltérés a predikált és mért értékek között

Logisztikus regresszió

Dichotóm kimeneti változó

Folytonos, dichotóm vagy ordinális prediktorok



A modell kicsit bonyolultabb:

$$p(Y_i) = \frac{1}{1 + e^{-(b0 + b1 * X_i)}}$$

Modell pontossága?

LL (log-likelihood) azaz a találat

Hasonlóság a predikált valószínűség és a mért kimenet között

-2LL azaz a tévedés

Eltérés a predikált valószínűség és a mért kimenet között

Statisztikai érték logisztikus regresszióban

- **Statisztikai érték**
  - A modellhez tartozó -2LL (tévedés) értéket összehasonlítjuk egy referenciamodell -2LL értékével, azaz tévedésével. A referenciamodellt baseline-nak szokás nevezni
    - Például azzal a modellel, ami nem tartalmaz prediktor változókat (ez felelne meg az ANOVA táblában található F érték értelmezésének lineáris regresszióban)
    - Vagy egy egyszerűbb, kevesebb prediktort tartalmazó modellel (ez felelne meg a Change Statistics tábla található F érték értelmezésének lineáris regresszióban)
    - **A -2LL értékek χ^2 eloszlást követnek**

$$\chi^2 = (-2LL_{baseline}) - \left(-2LL_{új}\right) = LL_{új} - LL_{baseline}$$

  - A tévedés csökkenéséhez (pontosság növekedésére) tartozó **χ^2 értékhez (ami egy statisztikai érték) hozzárendelhetjük a szignifikancia értéket**, ami segítségével megállapíthatjuk, hogy a modell szignifikánsan jobb-e, mint a baseline.

Statisztikai érték logisztikus regresszióban

- **Mi legyen a baseline?**
  - Lineáris regresszióban a kimeneti változó átlagát használtuk
  - Logisztikus regresszióban az átlagnak nem lenne értelme, hiszen valami vagy megtörtént, vagy sem, ezért **az előfordulási gyakoriságot fogjuk alapmodellnek használni**.
  - Például szeretnénk bejósolni, hogy egy terápiát valaki befejez-e vagy idő előtt megszakítja.
    - A vizsgált mintában 60 kliens van, ebből 42 befejezte a terápiát, 18 idő előtt megszakította.
    - Mire érdemes tippelnünk, ha semmit nem tudunk az emberekről, csak az előbbi arányt?

Befejezi a terápiát?		Jósolt kimenet: Nem	Jósolt kimenet: Igen
Tényleges kimenet	Nem	0	18
	Igen	0	42

- **Találatok**:
  - **Helyes találat**: 42 olyan ember van, akiről helyesen gondoltuk, hogy be fogja fejezni a terápiát
  - **Helyes elutasítás**: 0 olyan ember van, helyesen gondoltuk, hogy nem fogja befejezni a terápiát
- **Tévedések**:
  - **Téves riasztás**: 18 emberről helytelenül gondoltuk, hogy befejezi a terápiát
  - **Kihagyás**: 0 olyan ember van, aki befejezte a terápiát, még ha mi nem is erre számítottunk
- **Összesen**: a 60 emberből 42-őt találtunk el, így a **találati arány**: 42/60 = 70%

Statisztikai érték logisztikus regresszióban

- A következő modellbe emeljük be prediktorként azt, hogy mennyire együttműködő a kliens?
  - Minél együttműködőbb valaki, annál valószínűbb, hogy befejezi a terápiát. A nagyon alacsonyan együttműködők viszont valószínűleg meg fogják szakítani azt.
  - Az információ birtokában kicsit más jóslatot tehetünk:

		Jósolt kimenet	
		Nem	Igen
Tényleges kimenet	Nem	10	8
	Igen	4	38

- Találatok:
  - Helyes találat 38 és helyes elutasítás 10 ember
- Tévedések:
  - Téves riasztás 8 és Kihagyás: 4 ember
- Összesen: a 60 emberből 48-at találtunk el, így a találati arány: 48/60 = 80%

- Ebből kiszámolható az új és baseline modellhez tartozó -2LL (tévedés nagysága), és a változáshoz tartozó statisztikai érték: $\chi^2 = (-2LL_{baseline}) - (-2LL_{új}) = LL_{új} - LL_{baseline}$ illetve valószínűség.
- **Ha a χ^2 szignifikáns, akkor az együttműködést tartalmazó modell szignifikánsan jobban bejósolja a kimenetet, mint az egyszerű előfordulási gyakoriság.**
- **Ez felel meg a lineáris regresszió ANOVA táblájának**

Statisztikai érték logisztikus regresszióban

- A következő modellbe emeljük be prediktorként azt is, hogy volt-e már félbehagyott terápiája?
  - Ha valaki már szakított meg terápiát, kisebb valószínűséggel fogja ezt befejezni.
  - Az információ birtokában kicsit más jóslatot tehetünk:

		Jósolt kimenet	
		Nem	Igen
Tényleges kimenet	Nem	11	7
	Igen	4	38

- Találatok:
  - Helyes találat 38 és helyes elutasítás 11 ember
- Tévedések:
  - Téves riasztás 7 és Kihagyás: 4 ember
- Összesen: a 60 emberből 49-et találtunk el, így a találati arány: 49/60 = 81,66%

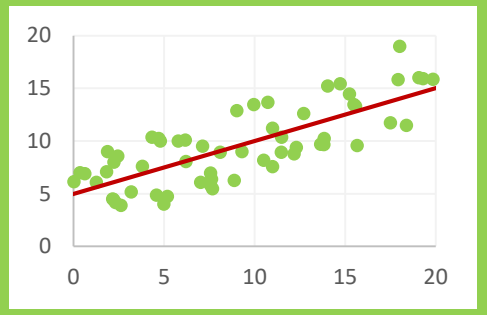
- Két χ^2 próbát is számíthatunk
  - megnézhetjük, hogy ez a modell, ami tartalmazza az együttműködést és a korábbi terápiáról az információt jobb-e az alapnál, ami csak a gyakoriságot használta (ez felel meg a lineáris regresszióban az ANOVA táblában a második modellhez tartozó F statisztikának)
  - Megnézhetjük, hogy ez a modell, ami tartalmazza az együttműködést és a korábbi terápiáról az információt jobb-e mint az eggyel egyszerűbb, ami csak a együttműködést tartalmazta (ez felel meg a lineáris regresszióban a Change Statistics-nek)

Eddigiek összegzése

Lineáris regresszió

Folytonos kimeneti változó

Folytonos vagy dichotóm prediktorok



A modell:

$Y_{pred,i} = b0 + b1 * X_i$

Modell pontossága?

SS_R, azaz a **legkisebb négyzetek módszer**ével számolt eltérés a predikált és mért értékek között

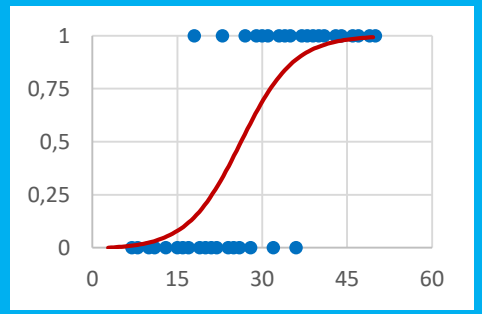
Modell statisztikai értéke: F

Modellek összehasonlítása: Change Statistics F-értéke

Logisztikus regresszió

Dichotóm kimeneti változó

Folytonos, dichotóm vagy ordinális prediktorok



A modell kicsit bonyolultabb:

$$p(Y_i) = \frac{1}{1+ e^{-(b0 + b1 * X_i)}}$$

Modell pontossága?

LL (log-likelihood) azaz a találat

Hasonlóság a predikált valószínűség és a mért kimenet között

-2LL azaz a tévedés

Eltérés a predikált valószínűség és a mért kimenet között

Modell statisztikai értéke: Modellhez tartozó χ^2 próba

Modellek összehasonlítása: Blockhoz tartozó χ^2 próba

Effect-size logisztikus regresszióban

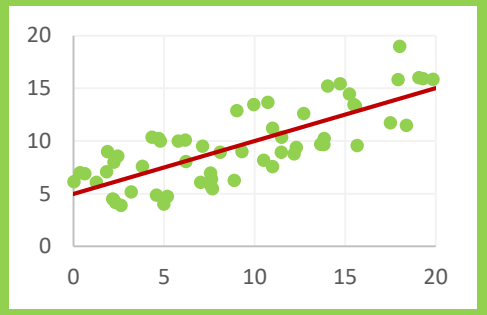
- Logisztikus regresszióban az R^2 értékét nem olyan egyszerű számolni, mint lineáris regresszióban
- R^2 helyett **pszeudo-R^2**-et használunk. *Pszeudo*, mert
  - bár nem a megmagyarázott varianciát adják meg (ez logisztikus regresszióban nem értelmezhető), de funkciójukban hasonlóak az R^2-hez, a hatás nagyságát mérik
  - Értékük 0%, ha nem a predikció pontatlanságát nem tudtuk csökkenteni
  - 100%, a kimenetet tökéletesen be tudjuk jósolni
- **Hosmer & Lemeshow R^2**
  - megadja, hogy a teljes pontatlanság mennyivel csökkent. SPSS alapból nem számolja
- **Cox & Snell R^2**
  - A Homer & Lemeshow R^2 elemszámra korrigált verziója
  - Túl szigorú, a hatás nagyságát alulbecsüli
- **Nagelkerke R^2**
  - A Cox & Snell R^2 szigorúságát korrigálja
  - Túl megengedő, a hatás nagyságát felülbecsli

Eddigiek összegzése

Lineáris regresszió

Folytonos kimeneti változó

Folytonos vagy dichotóm prediktorok



A modell:

$Y_{pred,i} = b0 + b1 * X_i$

Modell pontossága?

SS_R, azaz a **legkisebb négyzetek módszer**ével számolt eltérés a predikált és mért értékek között

Modell statisztikai értéke: F

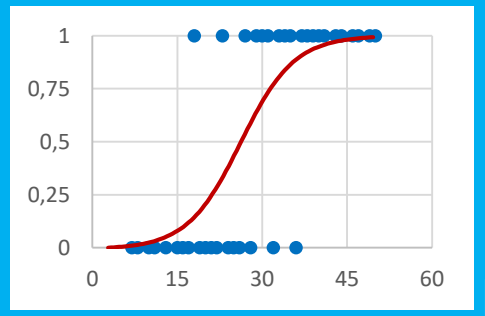
Modell effect-size mutatója: R^2

Modellek összehasonlítása: Change Statistics F-értéke

Logisztikus regresszió

Dichotóm kimeneti változó

Folytonos, dichotóm vagy ordinális prediktorok



A modell kicsit bonyolultabb:

$$p(Y_i) = \frac{1}{1+ e^{-(b0 + b1 * X_i)}}$$

Modell pontossága?

LL (log-likelihood) azaz a találat

Hasonlóság a predikált valószínűség és a mért kimenet között

-2LL azaz a tévedés

Eltérés a predikált valószínűség és a mért kimenet között

Modell statisztikai értéke: Modellhez tartozó χ^2 próba

Modell effect-size mutatója: Cox&Snell illetve Nagelkerke R^2

Modellek összehasonlítása: Blockhoz tartozó χ^2 próba

Együtthatók logisztikus regresszióban

- Bár bonyolultabb a képlet, a logisztikus regresszió is tartalmaz egy b0 konstanst és prediktoronként egy b1 meredekséget, melynek hasonló a jelentése a lineáris regresszióéhoz.

$$p(Y_i) = \frac{1}{1+ e^{-(b0 + b1 * X_i)}}$$

- **B érték**
  - Ha a b1 értéke pozitív, akkor a prediktor pozitív kapcsolatban van a kimeneti változóval, tehát ha nő a prediktor értéke, akkor valószínűbb lesz a vizsgált esemény bekövetkezése. Ha b1 értéke negatív, akkor negatív a prediktor és kimeneti változó közötti összefüggés. Vigyázat(!), a képletből is látszik, hogy az összefüggés nem lineáris!

- **Wald-teszt**
  - Hogy egy prediktor hatása szignifikánsan eltér-e a nullától lineáris regresszióban egy t-próbával ellenőriztük, logisztikus regresszióban a Wald-tesztet használjuk.
  - A Wald-teszt hajlamos a másodfajú hibára, tehát a prediktor hatását néha nem tudja kimutatni, ezért ha kaptunk egy általános képet az összefüggésekről általában újra futtatjuk a logisztikus regressziót a prediktorokat lépésenként a modellbe emelve, és a Change Statistics-ot elemezzük.

Együtthatók logisztikus regresszióban

- **Esélyhányados, odds-ratio, OR vagy Exp(B)**
  - A B értéket sokszor nehéz értelmezni, mert a kapcsolat logaritmikus. Helyette használható az esélyhányados.
  - Az esélyhányados értéke megadja, hogy egységnyi növekedés a prediktor változóban hányszorosára növeli a kimeneti változóval mért esemény bekövetkezésének valószínűségét.
  - **A prediktor változó egységnyi növekedése Exp(B)-vel szorozza a kimeneti változóval mért érték bekövetkezési valószínűségét**
  - Ha Exp(B) értéke 1-nél nagyobb, akkor pozitív a kapcsolat prediktor és kimeneti változó között.
  - Ha Exp(B) értéke 1-nél kisebb, akkor negatív a kapcsolat

- Példa: nézzük meg, mennyiben változtatja meg egy korábbi félbehagyott terápia a mostani terápia befejezésének esélyét! (az egyszerűség kedvéért legyen ez az egyetlen prediktor változó!)

		Befejezte ezt a terápiát?	
		Nem	Igen
Volt már félbehagyott terápia?	Nem volt	8	33
	Volt	10	9

Együtthatók logisztikus regresszióban

- Akiknél nem volt még félbehagyott terápia
  - 33-an fejezték be a terápiát, és 8-an szakították meg.
  - Itt a befejezés esélye $odds_{nemvolt} = p_{nemvolt} / q_{nemvolt} = 33 / 8 = 4.125$
- Akinél már volt félbehagyott terápia
  - 9-en fejezték be, és 10-en szakították meg.
  - Itt a befejezés esélye $odds_{volt} = p_{volt} / q_{volt} = 9 / 10 = 0.900$
- Az esélyhányados a korábbi megszakított és nem megszakított terápia esetén számolt esélyek hányadosa
  - $OR = odds_{volt} / odds_{nemvolt} = 0.900 / 4.125 = 0.218$
  - Azaz ha valakinek volt már korábban megszakított terápiája, akkor a mostani terápia befejezésének esélye egyötödére esik vissza!
- Logisztikus regresszióban OR számolható folytonos változókra is, de nehezebb értelmezni

		Befejezte ezt a terápiát?	
		Nem	Igen
Volt már félbehagyott terápia?	Nem volt	8	33
	Volt	10	9

Együtthatók logisztikus regresszióban

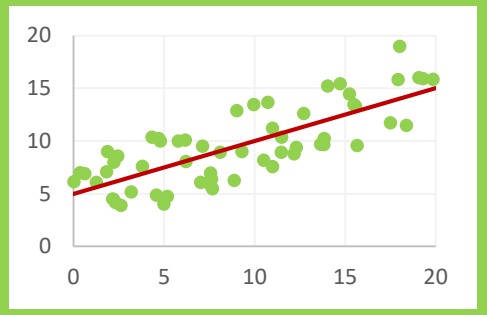
- Gyakoroljunk! Autizmus bejósolhatósága:
  - Nem:
    - Lányok körében 1 a 189-ből, fiúk körében 1 a 42-ből csecsemőt diagnosztizálnak ASD-vel
    - Mennyi az OR? Fogalmazd meg szavakkal, ez mit jelent?
  - Idősebb testvér ASD-vel
    - Prediktor: van = 1, nincs = 0
    - OR = 1,18
    - Fogalmazd meg szavakkal, ez mit jelent?
  - Apai és anyai életkor
    - 2 prediktor anyai kor és apai életkor: fiatalabb szülő (anya esetén 35 alatt, apa esetén 40 alatt) = 0, idősebb szülő = 1
    - OR_{anya} = 1.3 és OR_{apa} = 1.4
    - A logisztikus regresszióban OR számolható folytonos prediktorhoz is, de mivel nehezen értelmezhető, utólag sokszor dichotomizáljuk a prediktort, és úgy számolunk OR-t, mint itt is

Eddigiek összegzése

Lineáris regresszió

Folytonos kimeneti változó

Folytonos vagy dichotóm prediktorok



A modell:

$Y_{pred,i} = b0 + b1 * X_i$

Modell pontossága?

SS_R, azaz a **legkisebb négyzetek módszer**ével számolt eltérés a predikált és mért értékek között

Modell statisztikai értéke: F

Modell effect-size mutatója: R^2

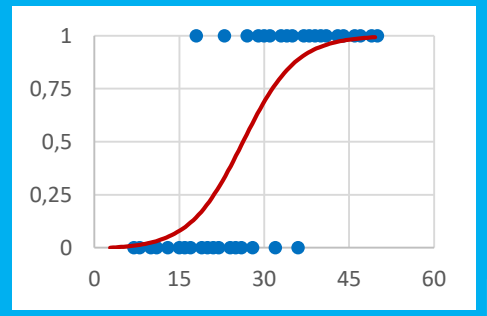
Modellek összehasonlítása: Change Statistics F-értéke

Prediktorok egyéni vizsgálata: koefficiensek és t-próbák

Logisztikus regresszió

Dichotóm kimeneti változó

Folytonos, dichotóm vagy ordinális prediktorok



A modell kicsit bonyolultabb:

$$p(Y_i) = \frac{1}{1+ e^{-(b0 + b1 * X_i)}}$$

Modell pontossága?

LL (log-likelihood) azaz a találat

Hasonlóság a predikált valószínűség és a mért kimenet között

-2LL azaz a tévedés

Eltérés a predikált valószínűség és a mért kimenet között

Modell statisztikai értéke: Modellhez tartozó χ^2 próba

Modell effect-size mutatója: Cox&Snell illetve Nagelkerke R^2

Modellek összehasonlítása: Blockhoz tartozó χ^2 próba

Prediktorok egyéni vizsgálata: koefficiensek és Wald-próbák

Feltételek

Feltételek

- **Változó típusa**
  - Kimeneti változó: dichotóm
  - Prediktor változó: Dichotóm, ordinális, skála típusú (néhány plusz beállítással nominális is lehet, mint ahogy a dummy változókkal lineáris regresszióba is be lehetett tenni nominális változót)
- **Nincs kollinearitás, multikollinearitás, külső változó**
  - A lineáris regresszióhoz hasonlóan itt sem jó, ha a változók túlzottan versengenek egymással – körültekintő tervezéssel, és a prediktorok korrelációs táblájának előzetes elemzésével ellenőrizhető
- **Linearitás és szóráshomogenitás**
  - a prediktor változók és a kimeneti változó között nem értelmezhető (hiszen a kimeneti változó dichotóm), ezért a kimeneti változó logaritmizált valószínűsége között kell fennálljon
- **Nincsenek (többdimenziós) outlierek**
  - Az egydimenziós outlierek az adattisztítás során, a többdimenziós outlierek a Cook távolsággal ellenőrizhetőek
- **Független mérések**

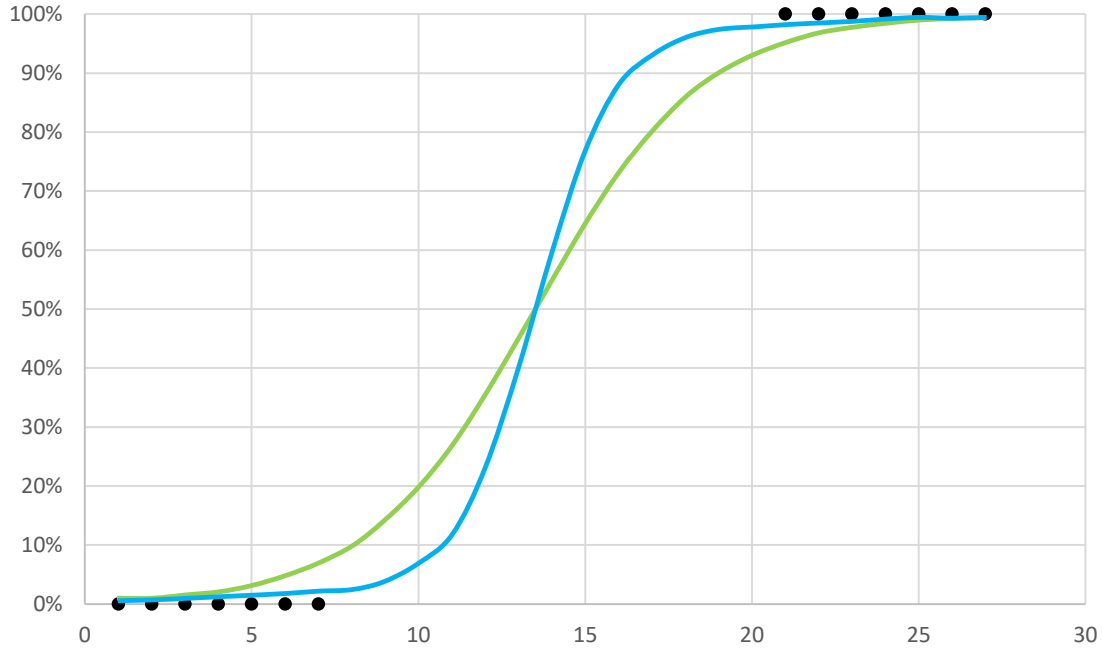
Feltételek

- **Nincsen olyan kombinációja a dichotóm prediktoroknak, amely nem fordul elő az adatbázisban**
  - A kombináció, amely egyáltalán nem fordul elő, nem bejósolható
  - Arról, ahol az anya idős, az apa fiatal, semmilyen információ nincs – nem bejósolható
  - Ilyen esetben a hibatagok aránytalanul megnőnek
  - Az elemzés előtt egy kereszttáblával ellenőrizni kell:
    - Ne legyen olyan kombináció, ami egyáltalán nem fordul elő
    - A kombinációk maximum 20%-a legyen N=5 alatt

Anyai kor	Apai kor	Autizmus
fiatal	fiatal	Nem
fiatal	fiatal	Igen
fiatal	fiatal	Nem
fiatal	fiatal	Nem
fiatal	idős	Igen
fiatal	idős	Igen
fiatal	idős	Nem
fiatal	idős	Nem
idős	idős	Igen
idős	idős	Igen
idős	idős	Nem
idős	idős	Igen

Feltételek

- **Nincs túlzott elkülönülés**
  - Nincs olyan prediktor változó (vagy prediktorok kombinációja), amely tökéletesen szétválasztja a kimeneti változó két csoportját
  - Ekkor több görbe is az adatokra illeszthető, és nem eldönthető, melyik jellemzi jól a kimeneti változóval mért esemény valószínűségét



Értelmezés

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	60	100,0
	Missing Cases	0	,0
	Total	60	100,0
Unselected Cases		0	,0
Total		60	100,0

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

Dependent Variable Encoding

Original Value	Internal Value
Nem	0
Igen	1

Nézzünk egy példát!
Szeretnénk megtudni, mitől függ, hogy a terápiát befejezi-e valaki, vagy idő előtt megszakítja (példa korábbról)

A kimeneti változó:

- Befejezte-e a terápiát (dichotóm változó)

Prediktor változók:

- Együttműködés mértéke (folytonos változó)
- Van-e korábbi megszakított terápiája (dichotóm)

A modellt úgy építettem fel, hogy először beléptettem az együttműködést majd egy második lépésben a korábbi megszakítást (tehát úgy, ahogy lineáris regresszióban a blockwise módszert tanultuk)
A minta 60 embert tartalmaz.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Befejezte-e a terápiát		Percentage Correct
			Nem	Igen	
Step 0	Befejezte-e a terápiát	Nem	0	18	,0
		Igen	0	42	100,0
	Overall Percentage				70,0

a. Constant is included in the model.

b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	,847	,282	9,046	1	,003	2,333

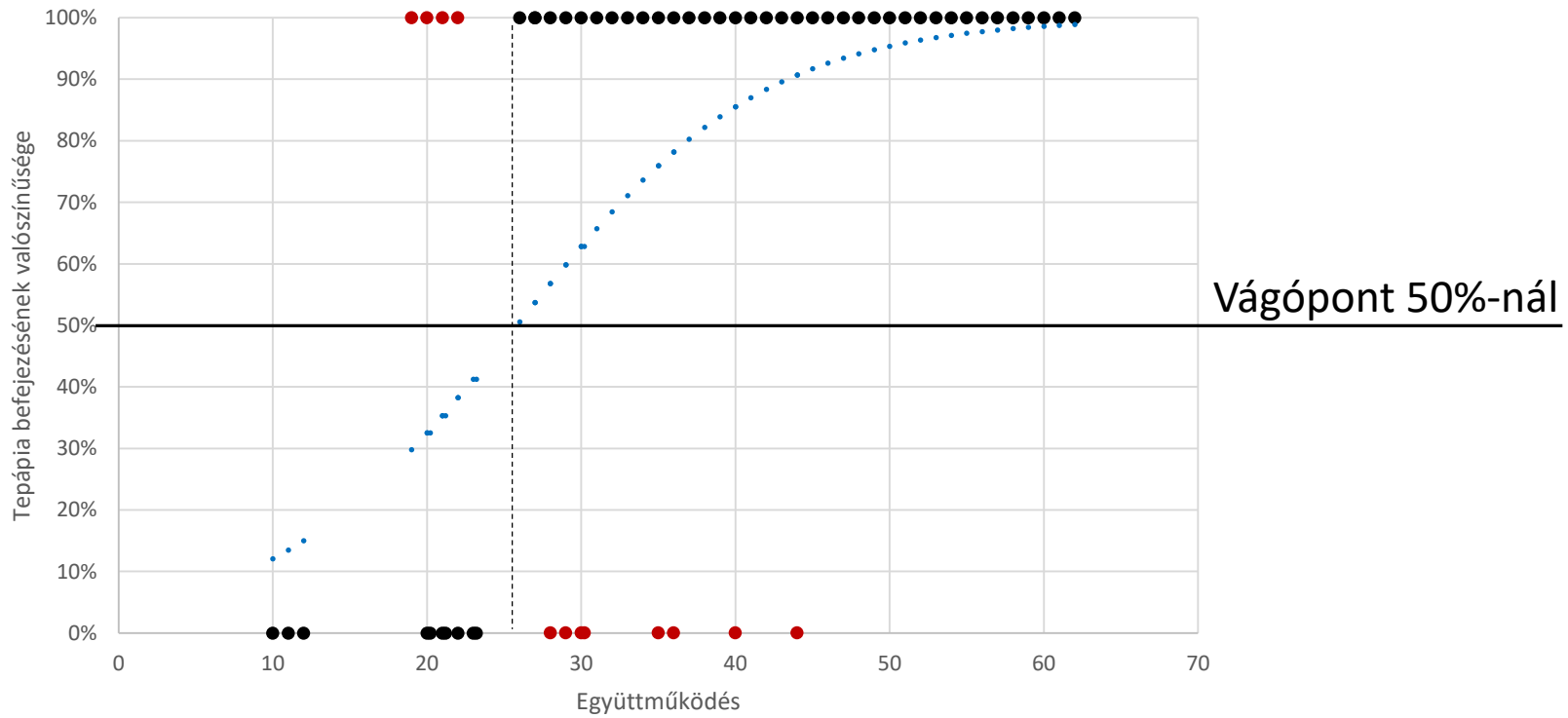
A Block 0 a prediktorok nélküli, alapmodellt tartalmazza. Mivel többen vannak, akik befejezték a terápiát, ezért mindenkit erre az oldalra predikálunk.

Block 1: Beginning Block

Classification Table[a]

Observed			Predicted		
			Befejezte-e a terápiát		Percentage Correct
			Nem	Igen	
Step 1	Befejezte-e a terápiát	Nem	10	8	55,6
		Igen	4	38	90,5
	Overall Percentage				80,0

a. The cut value is ,500



A Block 1 már tartalmaz egy folytonos prediktor változót, az együttműködés mértékét. Ennek segítségével felállítunk egy logisztikus modellt annak valószínűségének megállapítására, hogy valaki befejezi-e a terápiát vagy sem. Majd ennek eredményét a klasszifikációs táblában is megjelenítjük: akikhez a modell 50%-nál magasabb valószínűséget predikál a terápia befejezésére, azokat arra az oldalra soroljuk, míg azokat, akinél 50%-nál alacsonyabb valószínűséget látunk, azokat a terápiát nem befejezőkhöz predikáljuk.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	20,728	1	,000
	Block	20,728	1	,000
	Model	20,728	1	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	52,576[a]	,292	,414

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Még mindig a Block 1-et értelmezve:

Az Omnibus Test tábla Model sorában találjuk meg a X^2-próbát, mely az együttműködést, mint prediktort tartalmazó modellt hasonlítja össze a kezdeti, prediktort még nem tartalmazó alapmodellel (azaz a Block 0-val). Ez a lépés felel meg a lineáris regresszióban az ANOVA tábla ellenőrzésének. Megnézzük, a modell értelmezhető-e. Láthatjuk, hogy igen, a modell szignifikáns $X^2(1, N = 60) = 20.728$ $p < .001$

Majd a Model Summary táblából megnézzük, hogy milyen a modell magyarázóereje. A Cox-Snell és a Nagelkerke R^2 úgynevezett pszeudo R^2 értékek, csak hasonlítanak a hagyományos R^2-re, de értelmezésük azonos, azt jelölik, hogy a terápia befejezését /nem befejezését hány százalékban tudjuk magyarázni. Láthatjuk, hogy valahol 29.2 és 41.4% között mozog ez az érték.

Ha a modell szignifikáns, akkor meg lehet nézni, hogy a modellben szereplő prediktorok hogyan működnek. Ezt a Variables in the Equation táblában találjuk, ami megfelel a lineáris regresszió Coefficients táblájának.

A B oszlop tartalmazza a b0 és b1 értékeket. Az együttműködéshez tartozó meredekség pozitív, azaz az együttműködés növekedése megnöveli annak valószínűségét, hogy valaki sikeresen befejezi a terápiát (az értelmezésnél vigyázzunk, mert a kifejezés NEM lineáris!).

A Wald teszt felel meg a lineáris regresszió koefficiens táblájában látható t-próbának. Láthatjuk, hogy az együttműködés jó (szignifikáns) prediktora a terápia befejezésének. $W(1) = 12.114$ $p < .001$ A Wald teszt azonban nem elég pontos, érdemes az értelmezése mellett a prediktorok egyedi hatását úgy tesztelni, hogy blockwise építjük fel a modellünket, és a X^2-próbákkal vizsgáljuk minden prediktor hatását lépésenként.

Az Exp(B) az esélyhányadost tartalmazza. Megadja, hogy egységnyi változás az együttműködésben hogyan változtatja meg a terápia sikeres befejezésének esélyét. Mivel értéke 1 fölött van, így ebből is látjuk, hogy az együttműködés növekedésével nő a terápia befejezésének esélye.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	egyuttmuk	,125	,036	12,114	1	,001	1,134	1,056	1,217
	Constant	-3,238	1,125	8,284	1	,004	,039		

a. Variable(s) entered on step 1: egyuttmuk.

Block 2: Beginning Block

Classification Table[a]

Observed			Predicted		
			Befejezte-e a terápiát		Percentage Correct
			Nem	Igen	
Step 1	Befejezte-e a terápiát	Nem	11	7	61,1
		Igen	4	38	90,5
	Overall Percentage				81,7

a. The cut value is ,500

A Block 2-ben az együttműködés mellé belépett a modellbe a korábbi félbehagyott terápia is, így ez a modell már két prediktort tartalmaz.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4,990	1	,025
	Block	4,990	1	,025
	Model	25,718	2	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	47,586[a]	,349	,494

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

A predikciókat továbbra is a Classification Table-ben látjuk.

Az Omnibus Tests tábla Model sorából leolvasható, hogy ez a modell is szignifikáns, tehát értelmezhető $X^2(2, N = 60) = 25.718$ $p < .001$. A Block sor alapján pedig megállapíthatjuk, hogy szignifikánsan jobb, mint az előző, egyszerűbb, csak az együttműködést tartalmazó modell $X^2(1, N = 60) = 4.990$ $p = .025$. Tehát a Block sor megfelel a lineáris regresszió Change Statistics részének.

A Model Summary részből látjuk, hogy a hatásnagyság valahol 34.9 és 49.4% között van.

A Variables in the Equation táblából kiolvasható, hogy az együttműködés hatása szignifikáns $W(1) = 11.431$ $p = .001$, a hatás pozitív, mivel a meredekség (B) pozitív szám, illetve az esélyhányados (Exp(B) egynél nagyobb szám. A korábbi félbehagyott terápia hatása szintén szignifikáns $W(1) = 4.529$ $p = .033$, de negatív, tehát annak, akinek volt már félbehagyott terápiája, kisebb az esélye arra, hogy ezt befejezi. Az Exp(B) értéke 0.194, azaz kb. egyötödére esik vissza a terápia befejezésének esélye.
Továbbra is igaz: lehet nézni a Wald tesztet, de a korábbi félbehagyás hatását megbízhatóbban jelzi a Block sor X^2-próbája, mint a Wald teszt. A modell felépítésénél is érdemes akár többfajta prediktor-belépési sorrendet kipróbálni, és az eredményeket közösen értelmezni.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	egyuttmuk	,133	,039	11,431	1	,001	1,143	1,058	1,235
	felbehagyott	-1,640	,771	4,529	1	,033	,194	,043	,878
	Constant	-2,859	1,204	5,638	1	,018	,057		

a. Variable(s) entered on step 1: felbehagyott.