Logistická regrese

POLn4001

21.11.2024

Logistická regrese

- Technika, pomocí které se zjišťuje vliv nezávislých proměnných na závislou proměnnou

- Požadavky na proměnné:
  - Přesně jedna závislá proměnná – kategorické
  - Jedna nebo víc nezávislých proměnných
  - Nezávislé proměnné mohou být všech typů

- Důležitá je vždy teorie – cílem není počítat regresní modely s desítkami nezávislých proměnných

Logistická regrese

- Dokáže dát odpovědi na mnohé otázky
  - Zvyšuje se šance kandidáta na zvolení, pokud získá titul Mgr.?
  - Ovlivňuje šance Realu Madrid na výhru v zápase to, kdo je jeho aktuálním trenérem?
  - Mají studenti, kteří pravidelně navštěvují přednášky, vyšší šanci na úspěšné absolvování kurzu?
  - Mají uchazeči o práci s praxí vyšší šanci na úspěch ve výběrovém řízení?

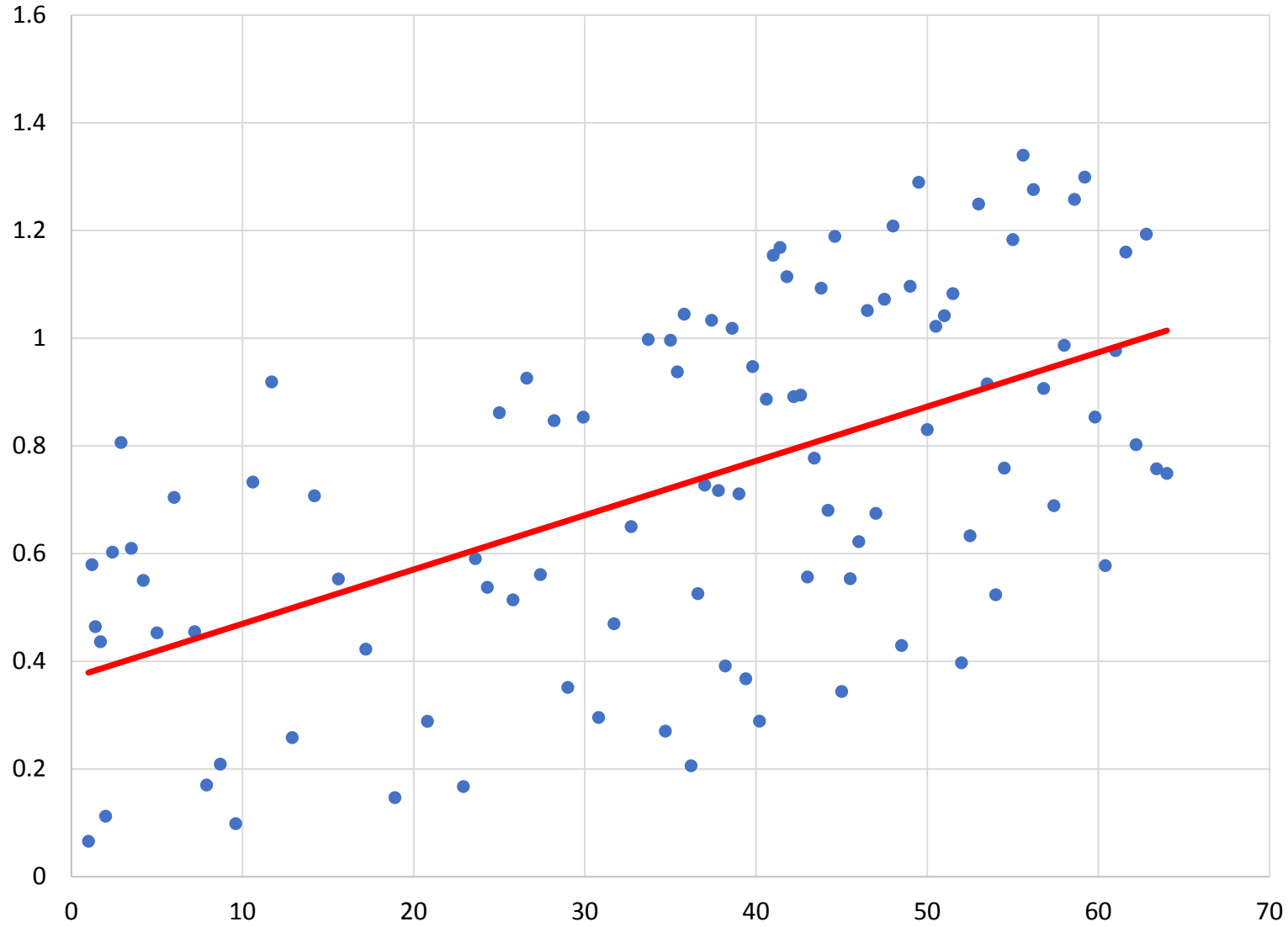
Logistická regrese – dva typy

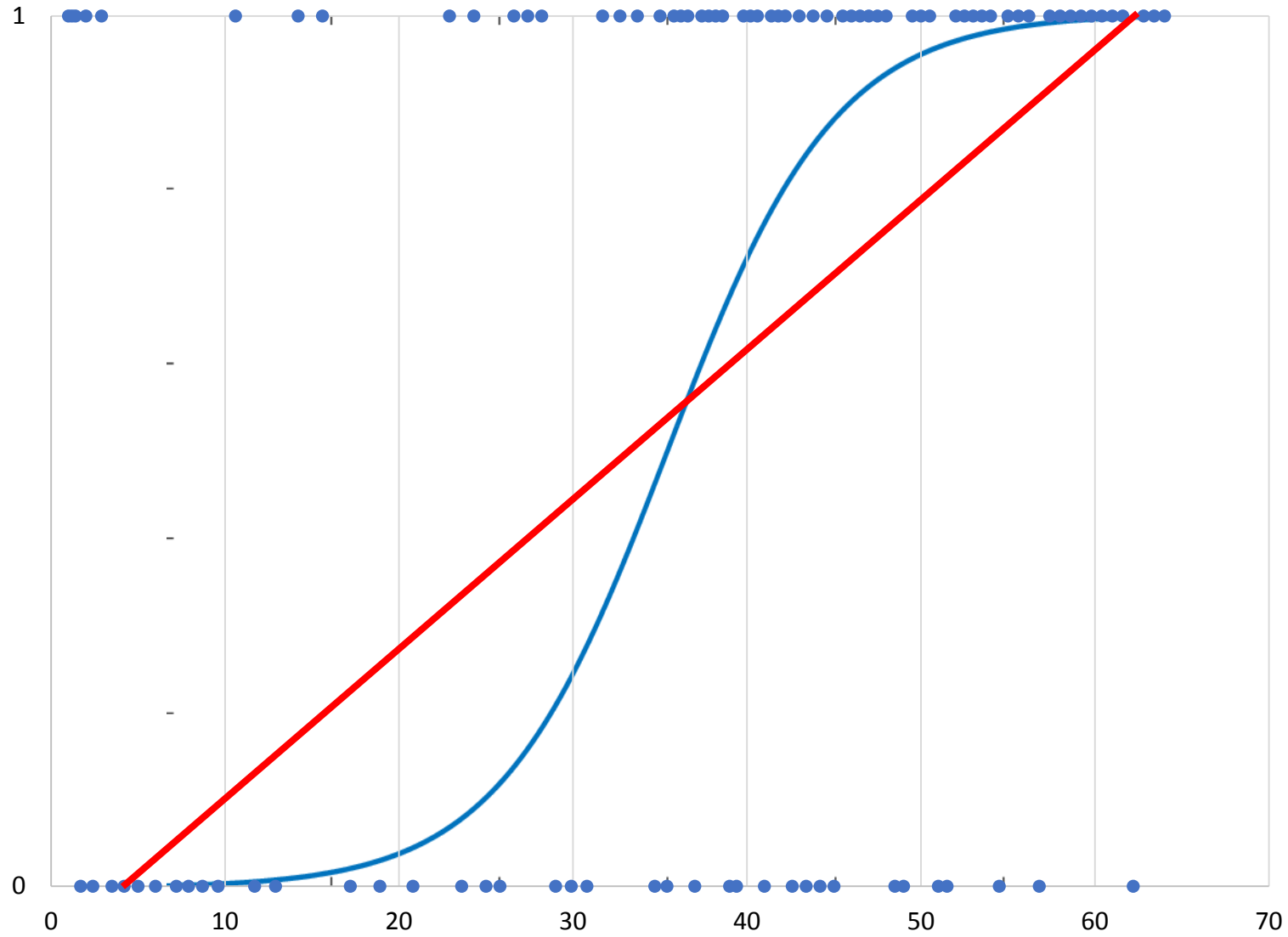
- Binární (binomial):
  - Závislá proměnná má dvě hodnoty (0/1)
  - Příklady – (0) Kandidát byl, (1) nebyl zvolený,
    (0) Volič se zúčastnil, (1) nezúčastnil voleb

- Multinomiální (multinomial, polynomial):
  - Závislá proměnná má více než dvě hodnoty (0/1/2)
  - Příklady – (0) Občan se nezúčastnil voleb, (1) zúčastnil a volil vládní stranu, (2) zúčastnil a volil opoziční stranu

Základní body

- Předpokladem lineární regrese je lineární vztah mezi nezávislými a závislou proměnnou
- Binární závislá proměnná toto neumožňuje, proto je tu lineární regrese nepoužitelná
- Logistická regrese absenci lineárního vztahu obchází použitím logaritmu


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70


1
0
0
10
20
30
40
50
60
70

Výstupy logistické regrese

- Co její pomocí můžeme zjistit?
  - Vhodnost modelu na analyzovaná data
  - Efekt každé nezávislé proměnné

- Důležité výstupy:
  - Log-likelihood
  - R^2
  - Konstanta
  - Odds ratio
  - Pravděpodobnosti

Log-likelihood

- Srovnává skutečná (pozorovaná) a modelem předpokládaná data
- Ukazuje, jak model pasuje na analyzovaná data
- Jeho hodnota vyjadřuje, jaký podíl variability zůstává po aplikaci modelu **nevysvětlený**
- Vyšší hodnoty ukazují na slabší sílu modelu a naopak

R^2

- V lineární regresi R^2 vyjadřuje, jaký podíl variability závislé proměnné je vysvětlen pomocí modelu
- V logistické regresi se R^2 interpretuje podobně, ale nejde o ekvivalent
- Více variant, SPSS produkuje Cox & Snell a Nagelkerke
- Mnozí autoři výpovědní hodnotu R^2 v logistické regresi zpochybňují

Konstanta a regresní koeficienty

- Konstanta:
  - Odhadovaná hodnota závislé proměnné, když je hodnota všech nezávislých proměnných rovna 0
  - Ve výstupu SPSS zapisováno jako Constant

- Koeficienty:
  - Odhadovaný efekt nezávislé proměnné na závislou proměnnou
  - Jak se změní hodnota závislé proměnné, pokud se hodnota nezávislé proměnné zvýší o jednotku
  - Náročnější intuitivní interpretace – hodnota je v podobě logaritmu
  - Ve výstupu SPSS zapisováno jako B

Odds ratio

- Ukazatel efektu prediktorů, jednoduchá interpretace
- Ukazuje, jak se se zvýšením nezávislé proměnné o jednotku mění šance na to, že nastane konkrétní výstup v závislé proměnné
- Hodnota 1 znamená žádný efekt, hodnoty nad 1 znamenají nárůst šancí, hodnoty pod 1 pokles šancí
- Ve výstupu SPSS zapisováno jako Exp(B)

Logistická vs. lineární regrese

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni}$$

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni})}}$$

Příklad

- Faktory úspěchu Zuzany Čaputové ve volbách 2019 na Slovensku

- Závislá proměnná:
  - Binární (0/1)
  - 1 = ZČ získala v obci nejvíc hlasů ze všech, 0 = ZČ nezískala tento počet hlasů

- Nezávislé proměnné:
  - Podíl obyvatel obce s VŠ vzděláním
  - Podíl Maďarů v obci
  - Podíl hlasů ĽSNS v parlamentních volbách
  - Velikost počtu obyvatel obce

Práce v SPSS

- Analyze → Regression → Binary Logistic
  - Závislá proměnná do *Dependent*
  - Nezávislé do *Covariates*

- Doporučené možnosti v *Options* a *Save* (Field, 281-282)

- Výběr metody:
  - Enter – všechny proměnné vstoupí do modelu okamžitě
  - Forward/Backward – postupné vkládání / ubírání
  - Závisí od cílů práce

Model 1 (VŠ, Maďaři, ĽSNS)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3160,702[a]	,263	,350

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VS	,148	,012	152,043	1	<,001	1,159
	Madari	,048	,003	190,001	1	<,001	1,049
	LSNS	-,093	,010	80,319	1	<,001	,911
	Constant	-,809	,169	22,824	1	<,001	,445

a. Variable(s) entered on step 1: VS, Madari, LSNS.

Interpretace efektů - VŠ

- Regresní koeficient B**:**
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní logaritmus hodnoty závislé proměnné
  - Zvýšení podílu lidí s VŠ vzděláním o 1 procentní bod vede k zvýšení logaritmu hodnoty závislé proměnné o 0,148 (ne příliš intuitivní)

- Poměr šancí (Odds Ratio):
  - Jednodušší interpretace efektu
  - 1,159 > 1 → zvýšení podílu lidí s VŠ vzděláním o 1 procentní bod **zvyšuje o 15,9 procenta šanci**, že ZČ získá v obcí nejvíc hlasů

Interpretace efektů – Maďaři

- Regresní koeficient B**:**
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní logaritmus hodnoty závislé proměnné
  - Zvýšení podílu Maďarů v obci o 1 procentní bod vede k zvýšení logaritmu hodnoty závislé proměnné o 0,048 (ne příliš intuitivní)

- Poměr šancí (Odds Ratio):
  - Jednodušší interpretace efektu
  - 1,049 > 1 → zvýšení podílu lidí s VŠ vzděláním o 1 procentní bod **zvyšuje o 4,9 procenta šanci**, že ZČ získá v obcí nejvíc hlasů

Interpretace efektů – ĽSNS

- Regresní koeficient B**:**
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní logaritmus hodnoty závislé proměnné
  - Zvýšení podílu hlasů pro ĽSNS v obci o 1 procentní bod vede k snížení logaritmu hodnoty závislé proměnné o 0,093 (ne příliš intuitivní)

- Poměr šancí (Odds Ratio):
  - Jednodušší interpretace efektu
  - 0,911 < 1 → zvýšení podílu hlasů pro ĽSNS v obci o 1 procentní bod **snižuje o 8,9 procenta šanci**, že ZČ získá v obci nejvíc hlasů

Logistická vs. lineární regrese

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni}$$

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni})}}$$

Variables in the Equation

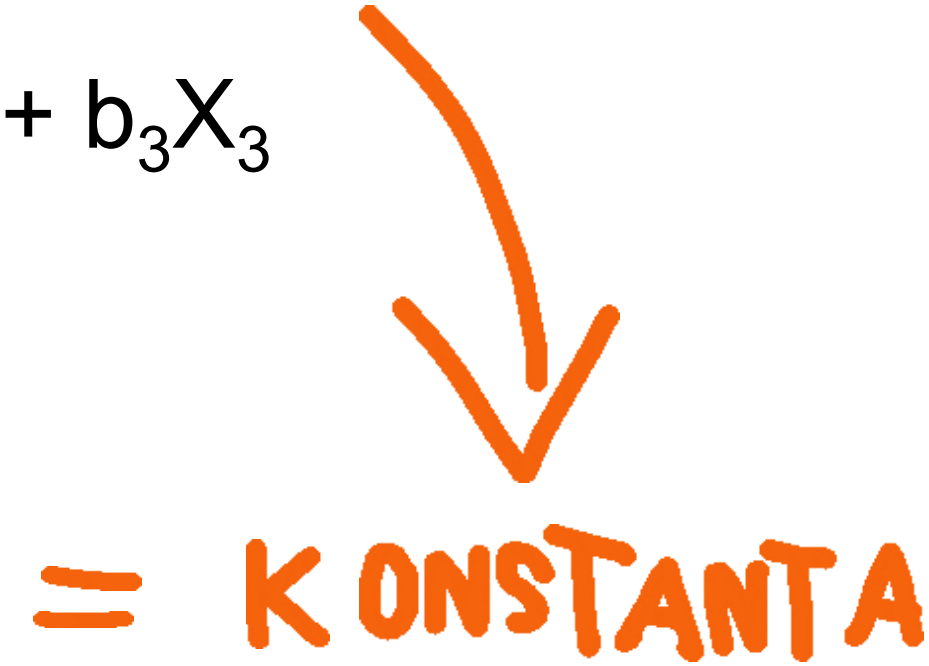
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VS	,148	,012	152,043	1	<,001	1,159
	Madari	,048	,003	190,001	1	<,001	1,049
	LSNS	-,093	,010	80,319	1	<,001	,911
	Constant	-,809	,169	22,824	1	<,001	,445

a. Variable(s) entered on step 1: VS, Madari, LSNS.

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i})}}$$

Modelová obec 1 – 0 % VŠ, 0 % Maďarů, 0 % ĽSNS

$$b_0 + b_1X_1 + b_2X_2 + b_3X_3$$



- $b_0 = -0{,}809$
- $b_1 = 0{,}148$
- $b_2 = 0{,}048$
- $b_3 = -0{,}093$
- $X_1 = 0$
- $X_2 = 0$
- $X_3 = 0$

-0,809 + 0,148*0 + 0,048*0 + (-0,093)*0

-0,809 + 0 + 0 + 0

= -0,809

- B → Exp(B)

- Exp(-0,809) = 0,445

- P = Exp(B) / (1 + Exp(B))
- P = 0,445 / (1 + 0,445)
- P = 0,445 / 1,445
- P = 0,3079

- Pravděpodobnost, že ZČ získá nejvíc hlasů ze všech kandidátů v obci s danými vlastnostmi je téměř **31 procent**

Modelová obec 2 – 24 % VŠ, 13 % Maďarů, 6,7 % ĽSNS

$$b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

- $b_0 = -0,809$
- $b_1 = 0,148$
- $b_2 = 0,048$
- $b_3 = -0,093$
- $X_1 = 24$
- $X_2 = 13$
- $X_3 = 6,7$

-0,809 + 0,148*24 + 0,048*13 + (-0,093)*6,7

-0,809 + 3,552 + 0,624 – 0,6231

= 2,7439

- B → Exp(B)

- Exp(2,7439) = 15,55

- P = Exp(B) / (1 + Exp(B))
- P = 15,55 / (1 + 15,55)
- P = 15,55 / 16,55
- P = 0,9396

- Pravděpodobnost, že ZČ získá nejvíc hlasů ze všech kandidátů v obci s danými vlastnostmi je 93,96, tedy téměř **94 procent**

Model 2 (VŠ, Maďaři, ĽSNS, Město)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3157,654[a]	,263	,351

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VS	,141	,012	128,476	1	<,001	1,152
	Madari	,048	,003	187,983	1	<,001	1,049
	LSNS	-,093	,010	80,482	1	<,001	,911
	Mesto	,377	,218	2,986	1	,084	1,458
	Constant	-,771	,171	20,401	1	<,001	,463

a. Variable(s) entered on step 1: VS, Madari, LSNS, Mesto.

Dummy proměnné

- Mají pouze dvě hodnoty (typicky 0/1)
- Nárůst jejich hodnoty „o jednotku“ je kompletně vyčerpá – není možný opakovaný nárůst jejich hodnoty
- Nižší hodnota (0) tak v modelu vystupuje v roli **referenční kategorie**, vůči které je efekt poměřován

Interpretace efektů – Město

- Regresní koeficient B:
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní logaritmus hodnoty závislé proměnné
  - Ve městech (1) je oproti malým obcím (0) logaritmus závislé proměnné vyšší o 0,377 (ne příliš intuitivní)

- Poměr šancí (Odds Ratio):
  - Jednodušší interpretace efektu
  - 1,458 > 1 → šance na to, že ZČ získá lokálně nejvíc hlasů, je ve městech ve srovnání s malými obcemi o 45,8 % vyšší
  - Jinými slovy, ve městech má ZČ podstatně vyšší šanci na vítězství než v malých obcích

Modelová obec 3 – 10 % VŠ, 5 % Maďarů, 8 % ĽSNS, Město (ano)

$$b_0 = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

- $b_0 = -0,771$
- $b_1 = 0,141$
- $b_2 = 0,048$
- $b_3 = -0,093$
- $b_4 = 0,377$
- $X_1 = 10$
- $X_2 = 5$
- $X_3 = 8$
- $X_4 = 1$

-0,771 + 0,141*10 + 0,048*5 + (-0,093)*8 + 0,377*1

-0,771 + 1,41 + 0,24 – 0,744 + 0,377

= 0,512

- B → Exp(B)

- Exp(0,512) = 1,67

- P = Exp(B) / (1 + Exp(B))
- P = 1,67 / (1 + 1,67)
- P = 1,67 / 2,67
- P = 0,6255

- Pravděpodobnost, že ZČ získá nejvíc hlasů ze všech kandidátů v obci s danými vlastnostmi je **62,6 procent**

Modelová obec 4 – 10 % VŠ, 5 % Maďarů, 8 % ĽSNS, Město (ne)

$$b_0 = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

- $b_0 = -0,771$
- $b_1 = 0,141$
- $b_2 = 0,048$
- $b_3 = -0,093$
- $b_4 = 0,377$
- $X_1 = 10$
- $X_2 = 5$
- $X_3 = 8$
- $X_4 = 0$

-0,771 + 0,141*10 + 0,048*5 + (-0,093)*8 + 0,377*0

-0,771 + 1,41 + 0,24 – 0,744 + 0

= 0,135

- B → Exp(B)

- Exp(0,135) = 1,14

- P = Exp(B) / (1 + Exp(B))
- P = 1,14 / (1 + 1,14)
- P = 1,14 / 2,14
- P = 0,5327

- Pravděpodobnost, že ZČ získá nejvíc hlasů ze všech kandidátů v obci s danými vlastnostmi je **53,3 procent**

Kategorické nezávislé proměnné

- Stejná logika jako u dummy proměnných
- Např. dny v týdnu, druhy zvířat, politické strany
- Postup:
  - Vytvořit dummy proměnné
  - Do modelu dát všechny kromě jedné – ta plní roli referenční kategorie
  - Koeficienty pro jednotlivé kategorie v modelu se poměřují vůči referenční kategorii
  - (*SPSS umožňuje místo tvorby dummy proměnných označit proměnnou jako kategorickou, výsledky modelu jsou stejné*)

Model 3 (VŠ, Maďaři, ĽSNS, velikostní kategorie obcí)

- Proměnná mapující počet obyvatel obce byla upravená na 4 dummy proměnné:
  - Obyv1 – obce do 500 obyvatel (1), ostatní obce (0)
  - Obyv2 – obce mezi 501 a 1000 ob. (1), ostatní obce (0)
  - Obyv3 – obce mezi 1001 a 5000 ob. (1), ostatní obce (0)
  - Obyv4 – obce s 5001+ ob. (1), ostatní obce (0)

- Jako referenční kategorie zvolena proměnná Obyv1
- Do modelu vstupují Obyv2, Obyv3 a Obyv4

ID	Kod_obce	Obyv1	Obyv2	Obyv3	Obyv4
82	503983	0	0	0	0
83	508209	0	0	0	0
84	508217	0	0	0	0
85	508276	0	0	0	0
86	508284	0	0	0	0
87	508292	0	0	0	0
88	508331	0	0	0	0
89	555509	0	0	0	0
90	501441	0	0	0	0
91	501450	0	0	0	0
92	501468	0	0	0	0
93	555517	0	0	0	0
94	501484	0	0	0	0
95	501492	0	0	0	0

Model 3 (VŠ, Maďaři, ĽSNS, velikostní kategorie obcí)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3109,823[a]	,275	,367

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VS	,129	,013	105,391	1	<,001	1,138
	Madari	,048	,004	186,666	1	<,001	1,049
	LSNS	-,095	,011	79,945	1	<,001	,909
	501-1000 ob	,351	,110	10,243	1	,001	1,421
	1001-5000 ob	,739	,107	47,345	1	<,001	2,094
	5001+ ob	,777	,226	11,801	1	<,001	2,174
	Constant	-,986	,179	30,442	1	<,001	,373

a. Variable(s) entered on step 1: VS, Madari, LSNS, 501-1000 ob, 1001-5000 ob, 5001+ ob.

Společná referenční kategorie Obyv1 (obce do 500 lidí)

Interpretace efektů – Vel. kategorie obcí

- Poměr šancí (Odds Ratio):
  - Obyv2: 1,421 > 1 → V obcích s 501-1000 obyvateli je **oproti obcím do 500 lidí** šance ZČ na lokální vítězství vyšší o 42,1 %
  - Obyv3: 2,094 > 1 → V obcích s 1001-5000 obyvateli je **oproti obcím do 500 lidí** šance ZČ na lokální vítězství téměř 2,1 násobně vyšší
  - Obyv4: 2,174 > 1 → V obcích s 5000 a víc obyvatel je **oproti obcím do 500 lidí** šance ZČ na lokální vítězství téměř 2,2 násobně vyšší

- Věcný závěr je, že v obcích s větší velikostí se ZČ dařilo podstatně lépe než v malých obcích

Důležité pro interpretaci efektů

- Vždy poznat vlastnosti nezávislých proměnných
  - Kardinální proměnné
  - Dummy proměnné
  - Kategorické proměnné s 3+ hodnotami

- Signifikantnost – závisí od povahy dat (vzorek/populace, reprezentativní vzorek?)

- Při počítání pravděpodobností nikdy nevynechat konstantu ani žádnou proměnnou, která je součástí modelu

Vizualizace výsledků

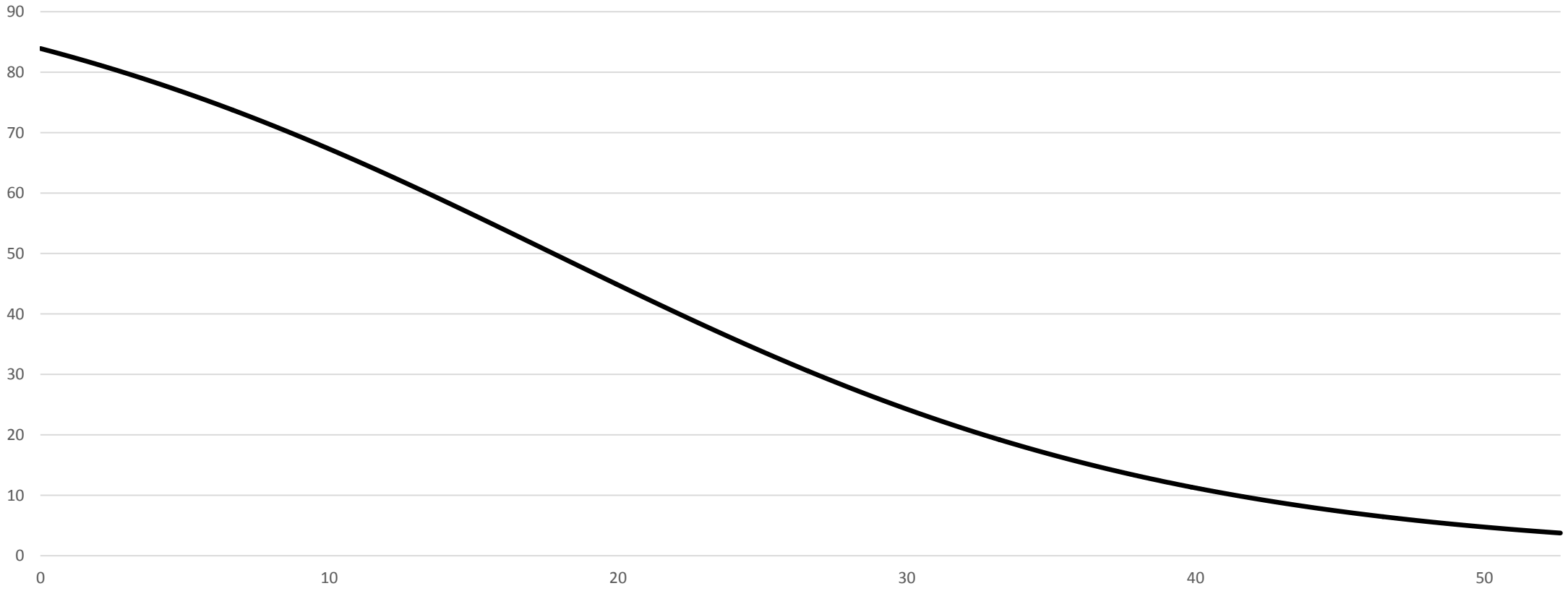
- V SPSS značně omezená

- (Limitovaná) možnost využití jiného softwaru:
  - Znázornění efektu jedné nezávislé proměnné
  - Je potřebné spočítat pravděpodobnosti s postupnou změnou dané nezávislé proměnné
  - Ostatní nezávislé proměnné musí být **po celý čas konstantní!**
  - Zanesení hodnot do grafu

Vizualizace výsledků (ĽSNS z Modelu 1)

VŠ	Maďaři	ĽSNS	P (ZČ Win)
15	5	0	83,90
15	5	0,1	83,78
15	5	0,2	83,65
15	5	0,3	83,52
15	5	0,4	83,39
15	5	0,5	83,26
15	5	0,6	83,13
15	5	0,7	83,00
15	5	0,8	82,87
…	…	…	…
15	5	52,63	3,76

Vizualizace výsledků (efekt L'SNS z Modelu 1)



Předpoklady a kontrola

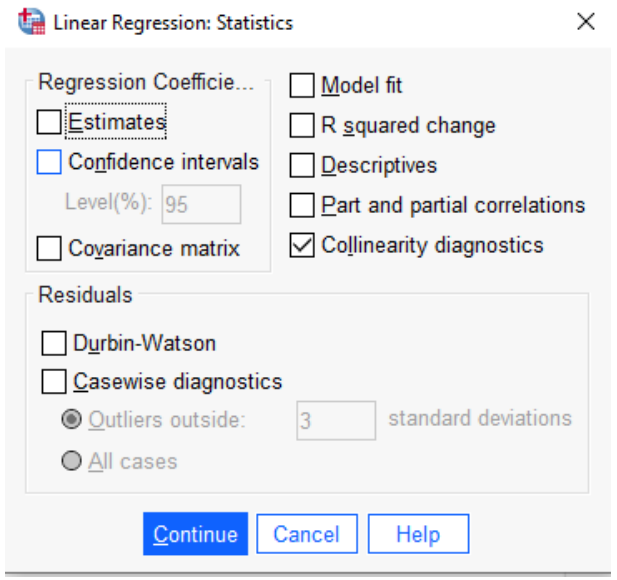
- Vhodný typ závislé proměnné
- Nezávislost pozorování
- Absence multikolinearity
- Rezidua

Testování multikolinearity

- Týká se pouze modelů s více než 1 nezávislou proměnnou
- Totožný postup jako u lineární regrese (SPSS nemá samostatné testování pro logistickou regresi)
- VIF – hodnoty nad 5 (10) indikují multikolinearitu
- Tolerance (1 / VIF) – hodnoty pod 0,1 (0,2) jsou problém
- Eigenvalues:
  - Proměnné by neměly mít vysokou variabilitu na stejných hladinách malých eigenvalues
- Pozor na dummy proměnné vytvořené z jedené kategorické proměnné

Testování multikolinearity

- Analyze → Regression – Linear
  - Nastavit proměnné
  - V *Statistics* zvolit *Collinearity Diagnostics*
  - Ostatní možnosti je možné vypnout (*Estimates*) – jde nám pouze o test multikolinearity



Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Penn State Worry Questionnaire	,575	1,741
	State Anxiety	,014	71,764
	Percentage of previous penalties scored	,014	70,479

a. Dependent Variable: Result of Penalty Kick

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Penn State Worry Questionnaire	State Anxiety	Percentage of previous penalties scored
1	1	3,434	1,000	,00	,01	,00	,00
	2	,492	2,641	,00	,04	,00	,00
	3	,073	6,871	,00	,95	,01	,00
	4	,001	81,303	1,00	,00	,99	,99

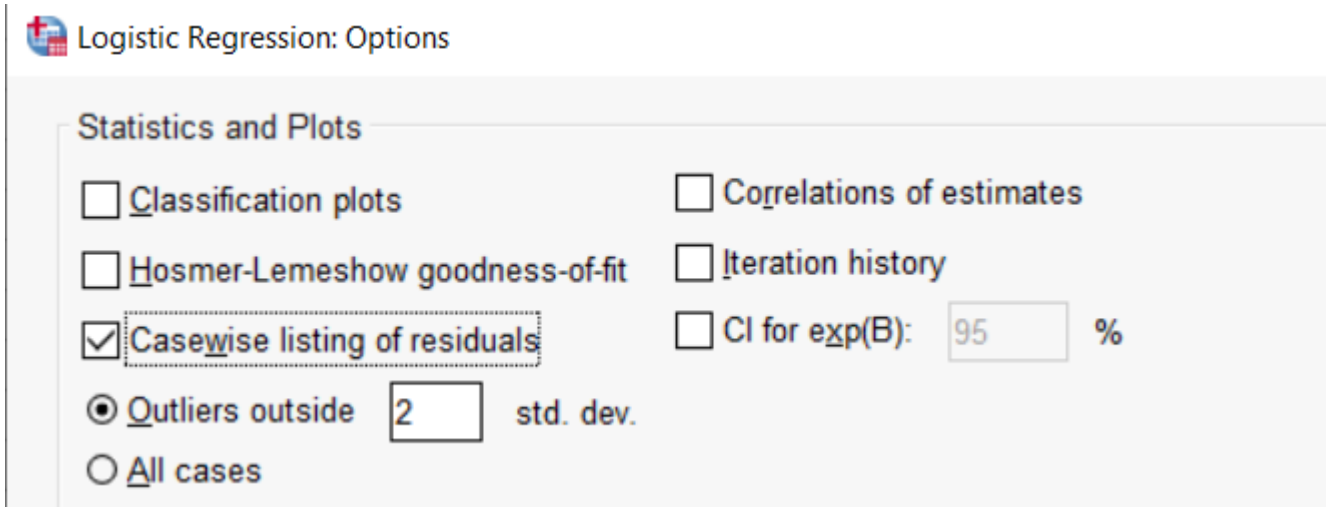
a. Dependent Variable: Result of Penalty Kick

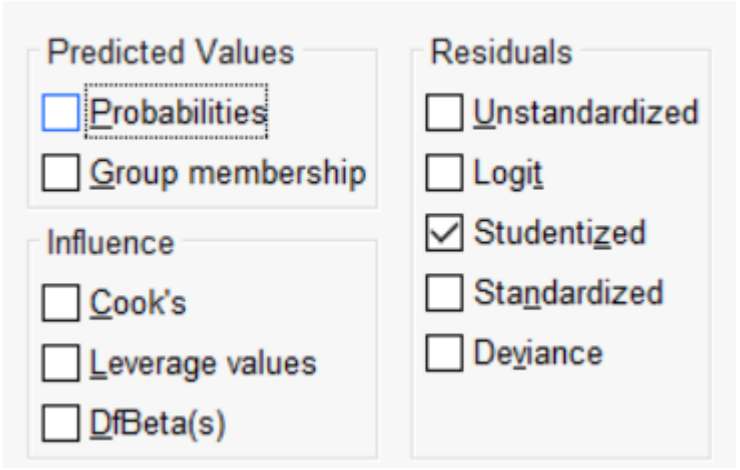
Testování multikolinearity

- Co v případě zjištění multikolinearity?
- Není možné zjistit unikátní efekty příslušných nezávislých proměnných
- Možnosti
  - Vyhodit jednu z příslušných proměnných
  - Separátní modely vždy pouze s jednou z daných proměnných

Rezidua

- Přijatelné hodnoty:
  - 95 % případů v rámci pásma -2 až 2
  - 99 % případů v rámci pásma -2,5 až 2,5

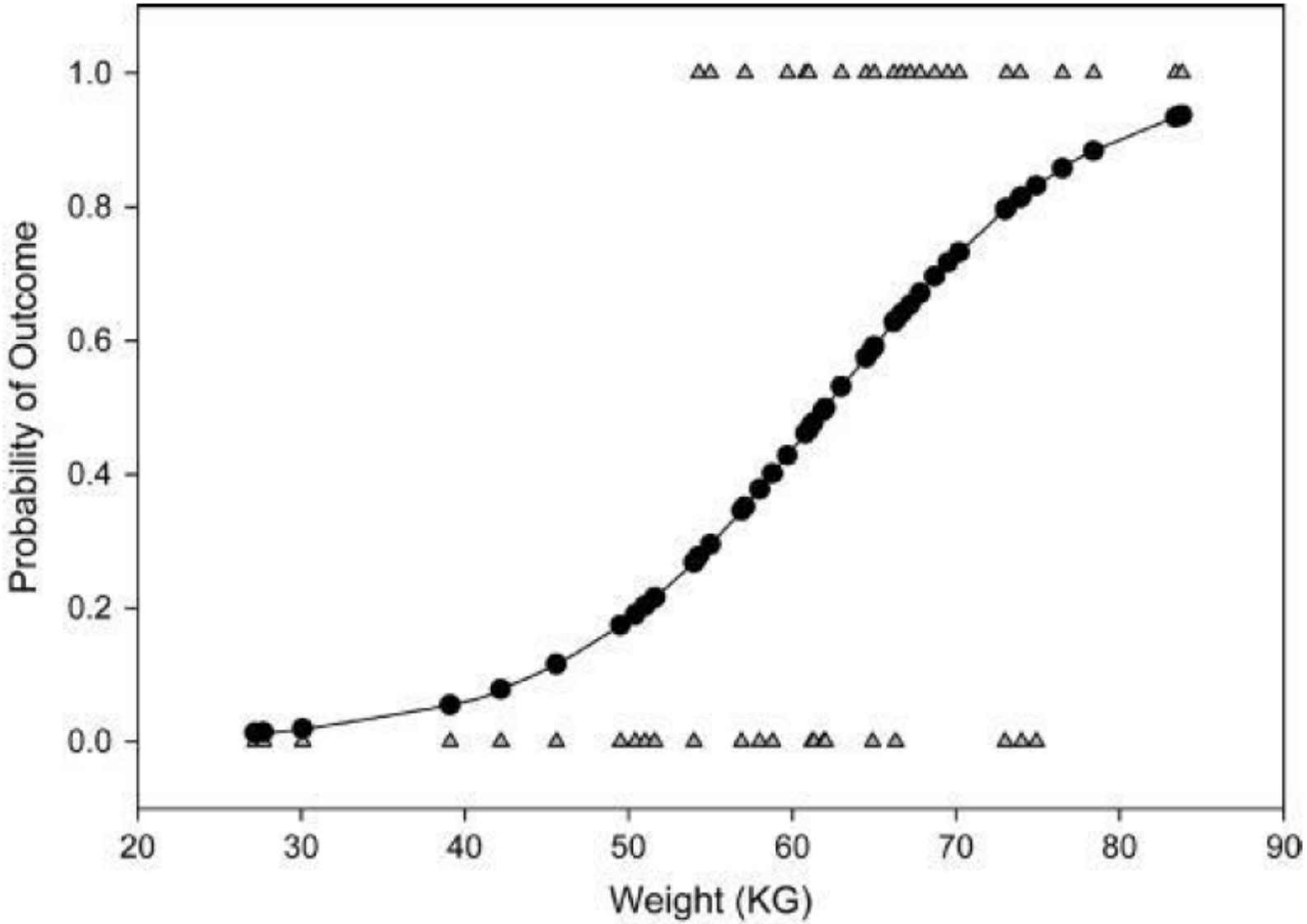


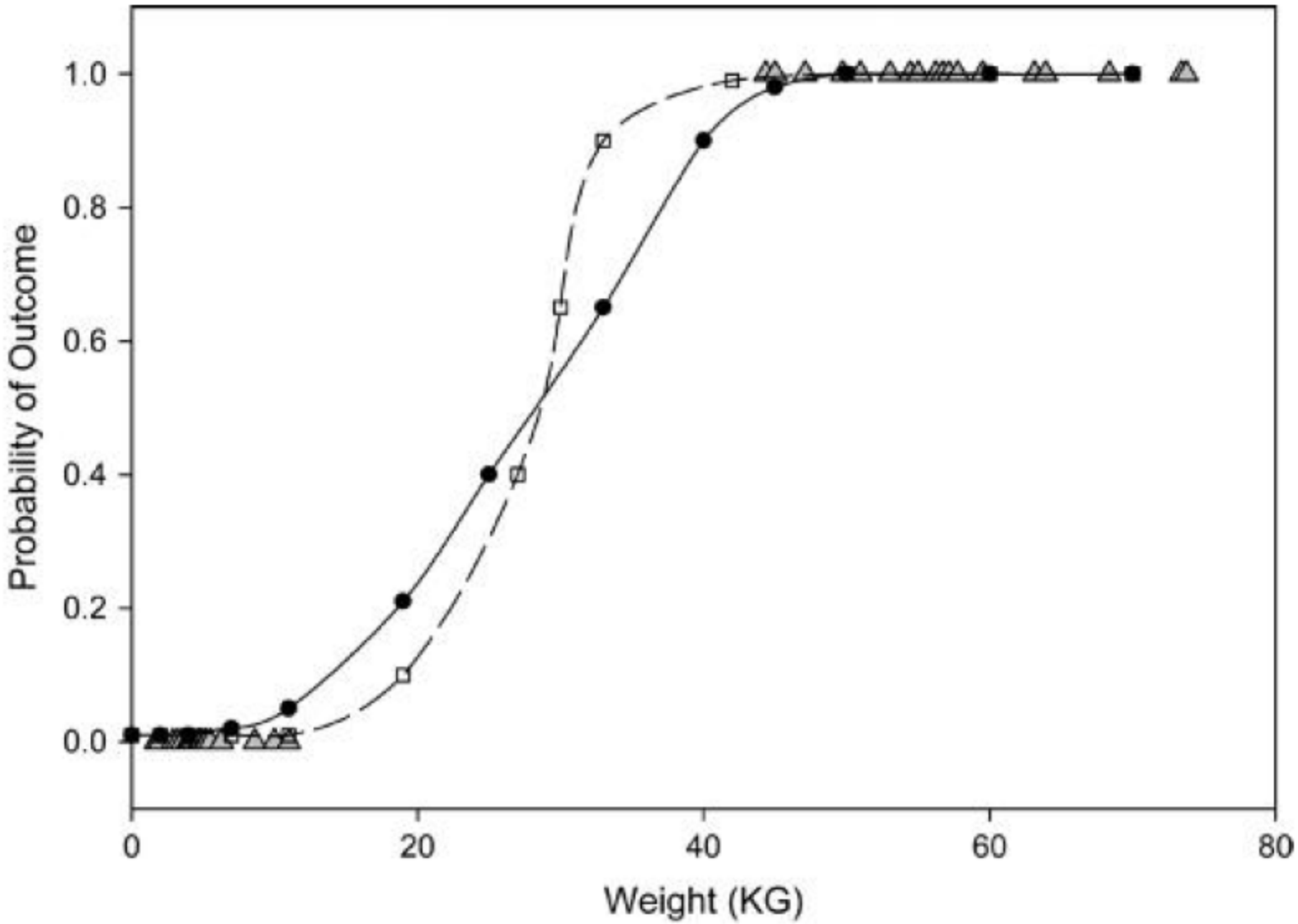


Možné problémy

- Nedostatek informací od prediktorů:
  - Neexistují data pro všechny kombinace hodnot proměnných
  - „Prázdná místa“ v kombinaci hodnot

- Kompletní oddělení:
  - Zdánlivý paradox - nastává, když pomocí nezávislé proměnné anebo proměnných dokážeme dokonale predikovat závislou proměnnou
  - Řešení – více dat anebo méně proměnných


Probability of Outcome
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
80
90
Weight (KG)


Probability of Outcome
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Weight (KG)