

# Logistická regrese

Peter Spáč

5.12.2019

# Logistická regrese

- Technika, pomocí které se zjišťuje vliv nezávislých proměnných na závislou proměnnou
- Využívá se v jiných případech než lineární regrese
- Rozdíl je v závislé proměnné:
  - Lineární – kardinální (anebo dlouhá ordinální)
  - Logistická – binární (0/1), krátká kategorická (0/1/2/3)
- Nezávislé proměnné mohou být všech typů

# Logistická vs. lineární regrese

- Lineární:
  - Zkoumá, jak se se změnou nezávislé proměnné o jednotku mění hodnota závislé proměnné
  - Např. jak se s počtem hodin strávených učením mění procentuální výsledek v testu
- Logistická:
  - Zkoumá, jak se se změnou nezávislé proměnné o jednotku mění **šance**, že nastane určitý výstup

# Logistická regrese



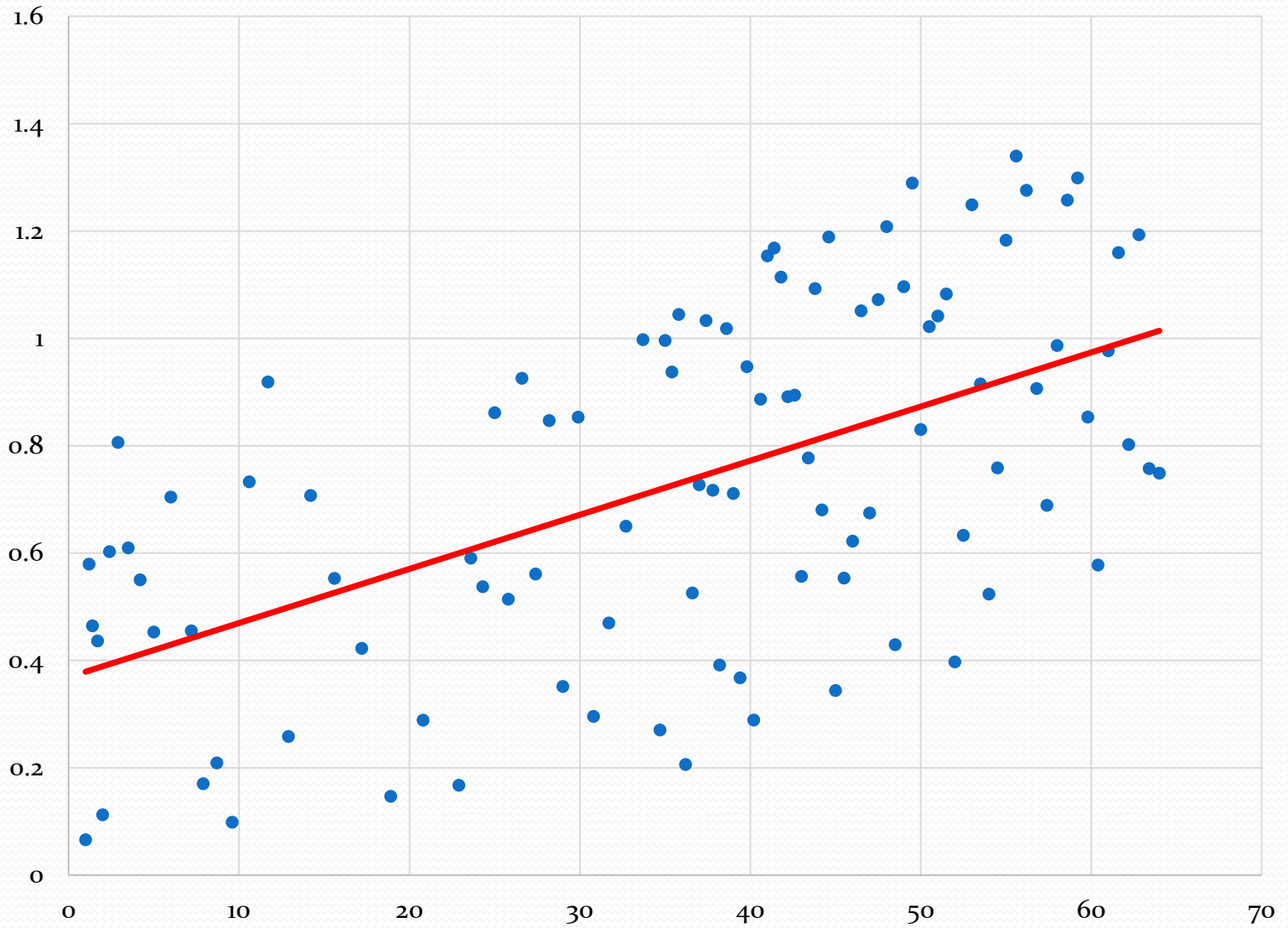
- Dokáže dát odpovědi na mnohé otázky
  - Zvyšuje se šance kandidáta na zvolení, pokud získá titul Mgr.?
  - Ovlivňuje šance Realu Madrid na výhru v zápase to, kdo je jeho aktuálním trenérem?
  - Mají studenti, kteří pravidelně navštěvují přednášky, vyšší šanci na úspěšné absolvování kurzu?
  - Mají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

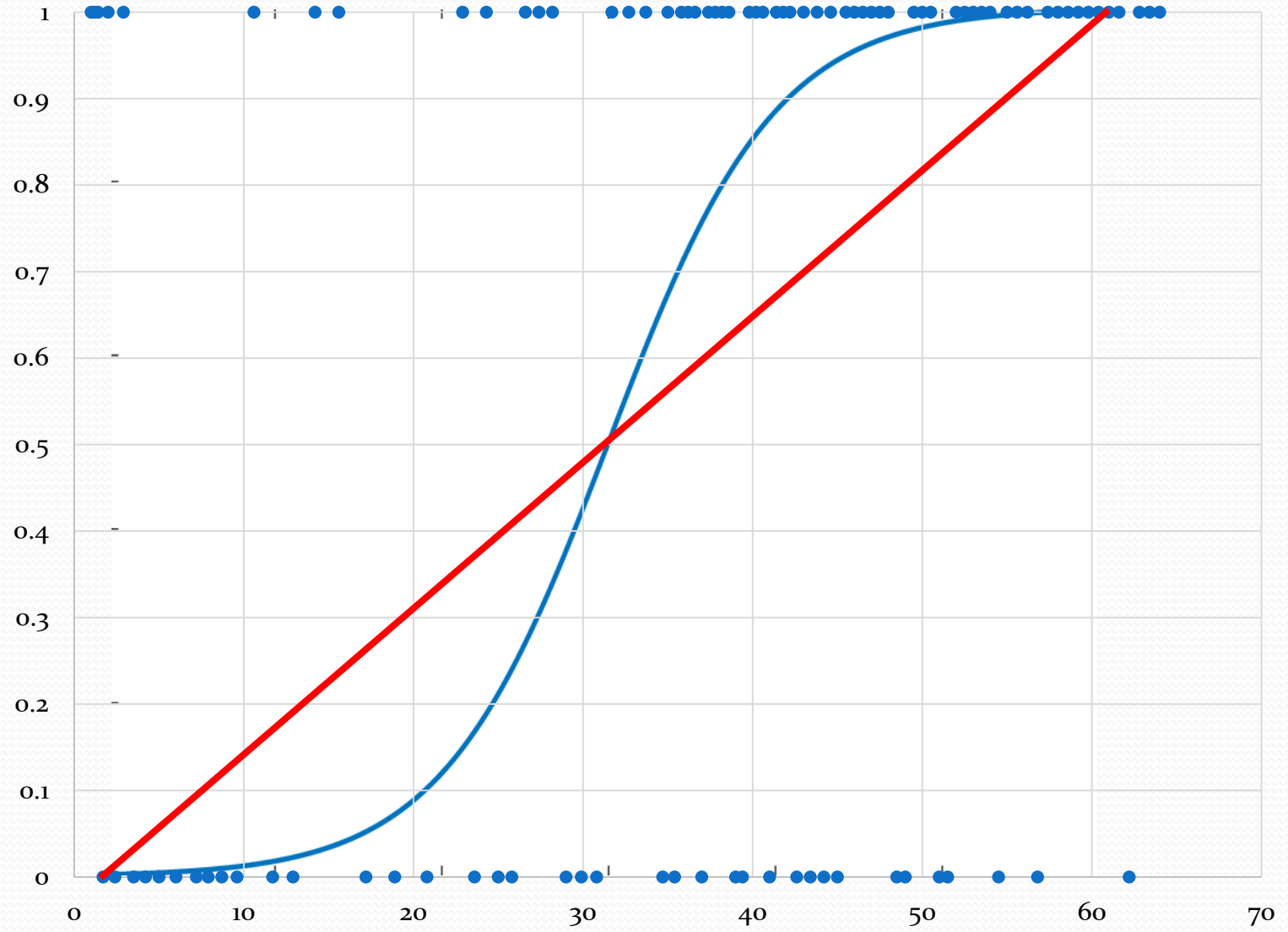
# Logistická regrese – dva typy

- Binární (binomial):
  - Závislá proměnná má dvě hodnoty (0/1)
  - Příklady – Kandidát byl/nebyl zvolený, volič se zúčastnil/nezúčastnil voleb
- Multinomiální (multinomial, polynomial):
  - Závislá proměnná má více než dvě hodnoty (0/1/2)
  - Příklady – Občan se nezúčastnil voleb / zúčastnil a volil vládní stranu / 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







# Logistická vs. lineární regrese

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

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

# Výstupy logistické regrese

- Co její pomocí můžeme zjistit?
  - Vhodnost modelu na analyzovaná data
  - Efekt každé nezávislé proměnné
- Důležité statistiky:
  - Log-likelihood
  - $R^2$
  - 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<sup>2</sup>

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

# Odds ratio

- Ukazatel efektu prediktorů, jednoduchá interpretace
- Ukazuje, jak se se změnou nezávislé proměnné o jednotku mění šance na to, že nastane konkrétní výstup v závislé proměnné
- Hodnoty nad 1 znamenají nárůst šancí, hodnoty pod 1 pokles
- Ve výstupu SPSS zapisováno jako  $\text{Exp}(B)$

# Předpoklady

1. Nezávislost pozorování

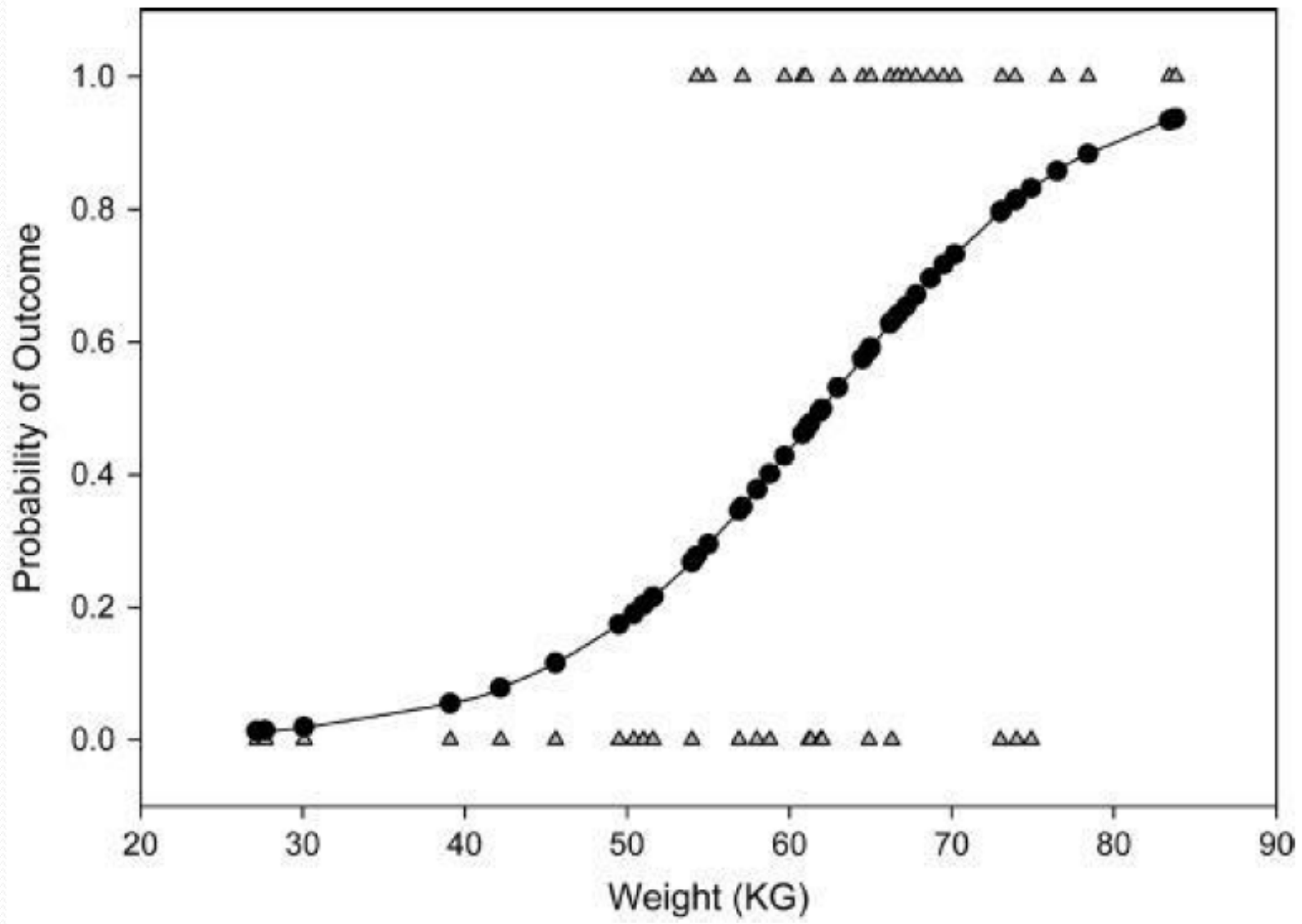
2. Absence multikolinearity

3. Linearita

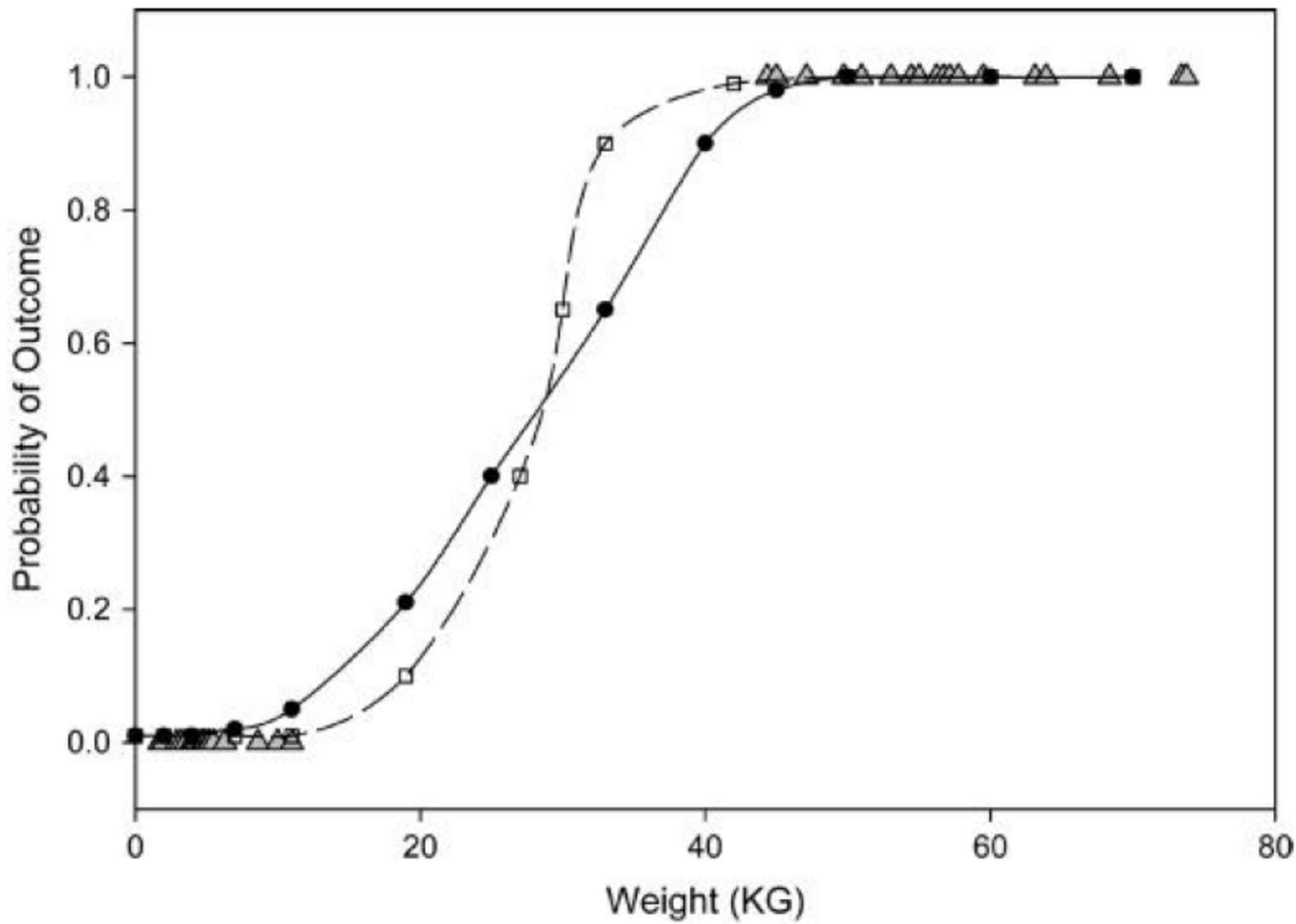
- V lineární regresi je podmínkou lineární vztah mezi nezávislou a závislou proměnnou
- V logistické toto neplatí – podmínkou je lineární vztah mezi **kontinuální** nezávislou proměnnou a logaritmem závislé proměnné

# 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







# Příklad

- Měření efektu léčebného zákroku na úspěšné vyléčení pacienta
- Závislá proměnná:
  - Binární (0/1)
  - Vyléčený/nevyléčený pacient
- Nezávislé proměnné:
  - Zákrok – realizovaný (1) vs. nerealizovaný (0)
  - Trvání nemoci ve dnech

# 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

# Práce v SPSS

- Zvolené možnosti:
  - Závislá - Cured
  - Nezávislé – Intervention, Duration
  - Metoda – Forward LR (nejdřív vytvoří model pouze s konstantou a následně bude přidávat prediktory v případě, že jsou přínosné)

### Iteration History<sup>a,b,c</sup>

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	154,084	,301
	2	154,084	,303
	3	154,084	,303

- a. Constant is included in the model.
- b. Initial -2 Log Likelihood: 154,084
- c. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

### Classification Table<sup>a,b</sup>

Observed			Predicted		
			Cured?		Percentage Correct
			Not Cured	Cured	
Step 0	Cured?	Not Cured	0	48	,0
		Cured	0	65	100,0
Overall Percentage					57,5

- a. Constant is included in the model.
- b. The cut value is ,500

# Výstupy

- Jako první je popsán model pouze s konstantou
- Uvedený je log-likelihood (154.08) – udává se v podobě -2 LL, což umožňuje posoudit signifikantnost
- Klasifikační tabulka
  - SPSS se při odhadu výstupu závislé proměnné přikloní k možnosti s vyšším počtem zastoupení (Cured)
  - Jde o nejlepší možný odhad, protože jiná data nemá
  - Správně tak zařadí 57,5 procent případů

### Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	,303	,190	2,538	1	,111	1,354

### Variables not in the Equation

	Score	df	Sig.
Step 0 Variables Intervention	9,771	1	,002
Duration	,609	1	,435
Overall Statistics	9,773	2	,008

# Výstupy

- Následuje sumarizace modelu s konstantou
- Podstatnější je seznam zatím nevložených proměnných
  - Signifikantní hodnota na konci (Score 9.773, sig. ,008) ukazuje, že vložení jedné nebo více těchto proměnných sílu modelu vylepší
  - Pokud by daná hodnota byla nesignifikantní, přidání jednotlivých proměnných by model neposílilo
- V dalším kroku tak SPSS přidá do modelu nejvhodnější proměnnou (Intervention)



### Iteration History<sup>a,b,c,d</sup>

Iteration		-2 Log likelihood	Coefficients	
			Constant	Intervention
Step 1	1	144,205	-,286	1,163
	2	144,158	-,288	1,228
	3	144,158	-,288	1,229

a. Method: Forward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 154,084

d. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

### Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9,926	1	,002
	Block	9,926	1	,002
	Model	9,926	1	,002

# Výstupy

- Vznikl nový model s přidáním *Intervention*
- Log-likelihood nového modelu = 144.16
  - Hodnota je nižší než u předešlého modelu (154.08)
  - Rozdíl obou je  $154.08 - 144.16 = 9.92$  a tento rozdíl je signifikantní
- Nový model je tak proti předešlému signifikantně lepší v předpovědi závislé proměnné

### Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144,158 <sup>a</sup>	,084	,113

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

### Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup> Intervention	1,229	,400	9,447	1	,002	3,417
Constant	-,288	,270	1,135	1	,287	,750

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

# Výstupy

- Údaj o Cox & Snell a Nagelreke  $R^2$  (mít na paměti výhradu o jejich výpovědní hodnotě)
- Klíčový výstup o efektech:
  - Regresní koeficient  $b$  – jak se při změně hodnoty nezávislé proměnné o jednotku změní logaritmus hodnoty závislé proměnné
  - Wald - ukazuje, zda koeficient  $b$  je signifikantně odlišný od nuly (v tomto případě je)

# Výstupy

- Odds ratio:
  - Jednoduchá interpretace efektu proměnné
  - $\text{Exp}(B) = 3.417 (> 1)$
  - Interpretace – pokud pacient podstoupí daný zákrok, jeho šance na vyléčení **se zvýší 3,4 násobně** proti těm pacientům, kteří zákrok nepodstoupí

**Model if Term Removed**

Variable	Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1 Intervention	-77,042	9,926	1	,002

**Variables not in the Equation**

	Score	df	Sig.
Step 1 Variables Duration	,002	1	,964
Overall Statistics	,002	1	,964

# Výstupy

- Výstup následku odebrání *Intervention* z modelu
  - Ukazuje, že odebráním proměnné by se významně změnila síla modelu → proměnná by se z modelu neměla odstranit
- Opět nabízí seznam nezařazených proměnných
  - (už bez *Intervention*, protože ta je v modelu)
  - Souhrnná hodnota Sig. ukazuje, že koeficient žádné z těchto proměnných není odlišný od nuly
  - Žádná další proměnná tak do modelu už nevstoupí
- Pokud by na začátku byla použita metoda **Enter**, všechny proměnné by do modelu vstoupily současně

### Iteration History<sup>a,b,c,d</sup>

Iteration		-2 Log likelihood	Coefficients		
			Constant	Intervention	Duration
Step 1	1	144,203	-,239	1,167	-,007
	2	144,156	-,235	1,233	-,008
	3	144,156	-,235	1,234	-,008

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 154,084

d. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

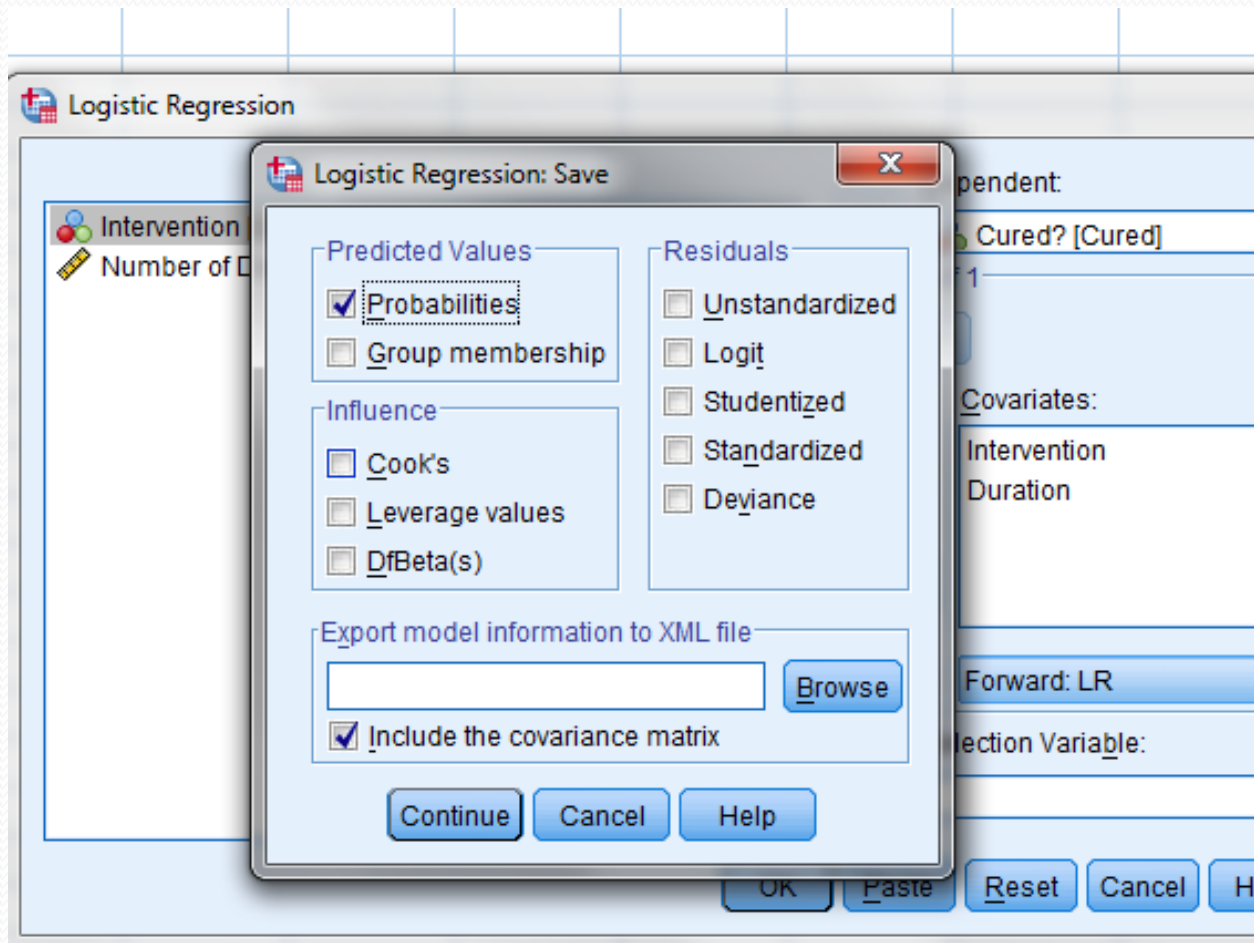
### Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Intervention	1,234	,415	8,854	1	,003	3,433
	Duration	-,008	,176	,002	1	,964	,992
	Constant	-,235	1,221	,037	1	,848	,791

a. Variable(s) entered on step 1: Intervention, Duration.



# Predicted probabilities



### Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Intervention	1,234	,415	8,854	1	,003	3,433
	Duration	-,008	,176	,002	1	,964	,992
	Constant	-,235	1,221	,037	1	,848	,791

a. Variable(s) entered on step 1: Intervention, Duration.

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

- Modelová osoba 1:

- Zákrok - 0 (ne)
- Počet dní - 7

$$b_0 + b_1 X_{1i} + b_2 X_{2i}$$

- $b_0 = -0,235$
- $b_1 = 1,234$
- $b_2 = -0,008$
- $X_1 = 0$
- $X_2 = 5$

$$-0,235 + 1,234 * 0 + (-0,008 * 5)$$

$$-0,235 + 0 - 0,04$$

$$-0,275$$

- $B \rightarrow \text{Exp}(B)$
- $\text{Exp}(-0,275) = 0,7596$
- $P = \text{Exp}(B) / (1 + \text{Exp}(B))$
- $P = 0,7596 / (1 + 0,7596)$
- $P = 0,7596 / 1,7596$
- $P = 0,4316$
- Pravděpodobnost vyléčení osoby, která nepodstoupila zákrok a jejíž nemoc trvala 5 dní je **43 procent**
- Co osoba, která podstoupila zákrok s dobou trvání nemoci 10 dní?

$$b_0 + b_1 X_{1i} + b_2 X_{2i}$$

- $b_0 = -0,235$
- $b_1 = 1,234$
- $b_2 = -0,008$
- $X_1 = 1$
- $X_2 = 10$

$$-0,235 + 1,234 * 1 + (-0,008 * 10)$$

$$-0,235 + 1,234 - 0,08$$

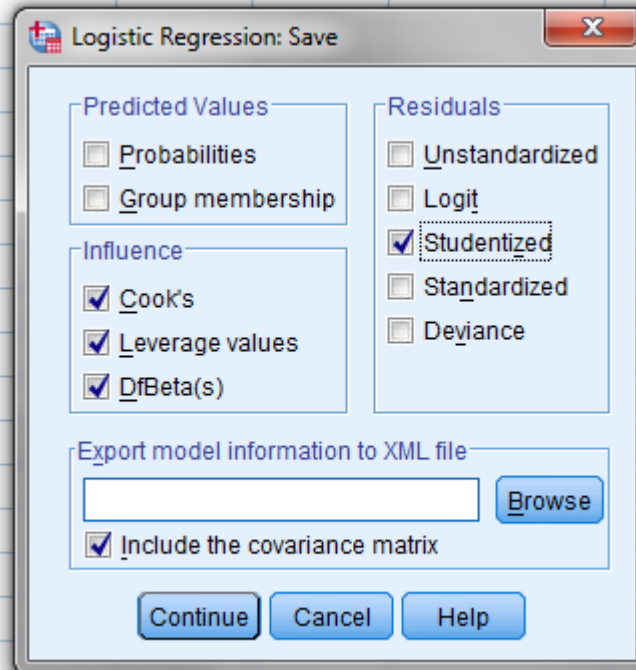
$$0,919$$

- $B \rightarrow \text{Exp}(B)$
- $\text{Exp}(0,919) = 2,5068$
- $P = \text{Exp}(B) / (1 + \text{Exp}(B))$
- $P = 2,5068 / (1 + 2,5068)$
- $P = 2,5068 / 3,5068$
- $P = 0,7148$
- Pravděpodobnost vyléčení osoby, která podstoupila zákrok a jejíž nemoc trvala 10 dní je **71 procent**

# Následná kontrola

- Dva základní cíle
- 1. Zjistit, pro která data není model vhodný
  - Identifikace odlehlých případů (outliers)
  - Rezidua
- 2. Zjistit, které případy mají nadměrný vliv na model:
  - Cookova vzdálenost (Cook's distance)
  - DFBeta
  - Leverage

- Analyze → Regression → Binary Logistic
  - Položka *Save*
- SPSS požadované hodnoty uloží do automaticky vytvořených proměnných
  - COO\_1, LEV\_1, SRE\_1, DFBo\_1, DFB1\_1, ...





# Přijatelné hodnoty

- Cook's distance a  $DFBeta < 1$
- Rezidua:
  - 95 % případů v rámci pásma -2 až 2
  - 99 % případů v rámci pásma -2,5 až 2,5
  - Studentized jsou vhodnější než standardized
- Leverage:
  - $(k + 1) / N \rightarrow$  počet nez. proměnných zvýšených o jedna se vydělí počtem případů
  - Přijatelné hodnoty jsou do 2 až 3 násobku takto spočítané hodnoty

# Následná kontrola

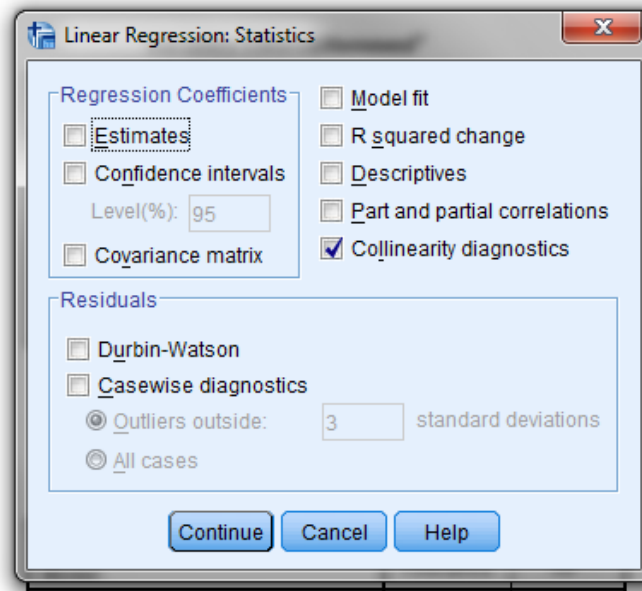
- Skutečný problém vytvářejí až případy, jenž mají vysoká rezidua (nad 2 až 3) a **současně** nadměrnou leverage
- Jejich negativní efekt je silnější v malých vzorcích (je zde méně případů, které jejich efekt vyváží)
- Pro odstranění případů z analýzy je potřebné mít dobrý důvod

# 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 ( $1 / \text{VIF}$ ) – hodnoty pod 0,1 (0,2) jsou problém
- Eigenvalues:
  - Proměnné by neměly mít vysokou variabilitu na stejných hladinách malých eigenvalues

# Testování multikolinearity

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



### Coefficients<sup>a</sup>

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<sup>a</sup>

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

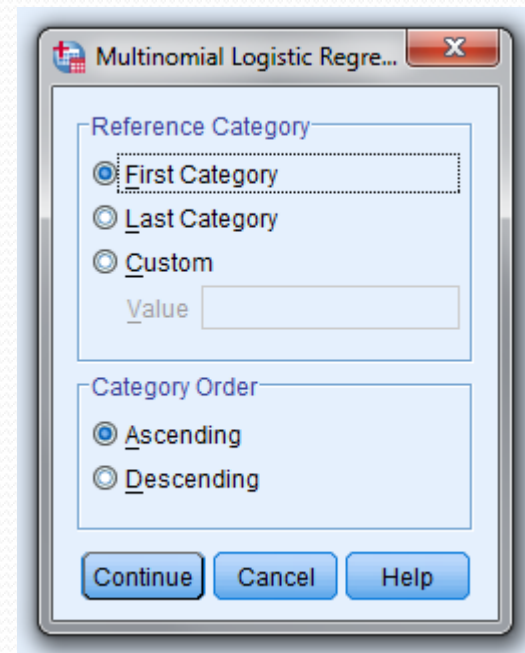
- Co v případě zjištění multikolinearity?
- Není možné zjistit unikátní efekty příslušných nezávislých proměnných
- Možnosti
  - Vyhodit jednu z příslušných proměnných
  - Separátní modely vždy pouze s jednou z daných proměnných

# Multinomiální logistická regrese

- Od binární se odlišuje pouze povahou závislé proměnné
- Závislá proměnná má více než dvě hodnoty (0/1/2/3)
- Postup je úplně totožný jako u binární log. regrese
- Výsledky se interpretují vždy k jedné z hodnot závislé proměnné, jež je stanovena jako referenční (jako nula u binární log. regrese)

# Práce v SPSS

- Analyze → Regression → Multinomial Logistic
  - Závislá proměnná do *Dependent* (v Reference category vybrat referenční kategorii)
  - Nezávislé do *Covariates*





# Výstupy

- Výstupy se interpretují ve vztahu k referenční kategorii
- Pokud z 0/1/2 je nula referenční, tak:
  - 1 VS. 0
  - 2 VS. 0

Parameter Estimates

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Success of Chat-Up Line <sup>a</sup>	Intercept	-1,153	,336	11,802	1	,001			
	Funny	,135	,048	7,877	1	,005	1,145	1,258	
	Good_Mate	,137	,051	7,220	1	,007	1,146	1,266	
Go Home with Person	Intercept	-3,668	,521	49,555	1	,000			
	Funny	,383	,070	29,630	1	,000	1,467	1,684	
	Good_Mate	,149	,077	3,772	1	,052	1,160	1,349	

a. The reference category is: No response/Walk Off.