

PSYb2520

Statistická analýza dat v psychologii II

Přednáška 4

Logistická regrese

Logistic regression

Předpovídáme pohlaví pachatele

Víme, že pachatel nosí náušnici/e a napsal dopis se skórem emočních adjektiv 8.

Víme, že...

- náušnice nosí 21% mužů a 85% žen
- na škále přítomnosti emočních adjektiv od 1 do 13 mají ženy průměr 9,1 a muži pouze 4,5.

Jaká je pravděpodobnost, že pachatel je žena?

Nejprve využijme informaci o náušnici

- náušnice nosí 23% mužů a 85% žen
 - $P(\text{nosí}|\text{žena})=85\%$ a $P(\text{nosí}|\text{muž})=23\%$
 - Jenže my víme, že nosí a potřebujeme
pravděpodobnost pohlaví – $P(\text{žena}|\text{nosí})=?$
 - $$\begin{aligned} P(\text{ž}|\text{n}) &= P(\text{n}|\text{ž})P(\text{ž})/P(\text{n}) = \\ &= P(\text{n}|\text{ž})P(\text{ž})/(P(\text{n}|\text{ž})P(\text{ž})+P(\text{n}|\text{m})P(\text{m})) = \\ &= 0,85*0,5/(0,85*0,5+0,23*0,5) = \mathbf{0,79} \end{aligned}$$
-

CROSSTABS

/TABLES=pohlavi BY nausnice

/CELLS=COUNT ROW

/COUNT ROUND CELL.

pohlavi * nausnice Crosstabulation

		nausnice			
		0 nenosí	1 nosí	Total	
pohlavi	0 mužské	Count	10	3	13
		% within pohlavi	76,9%	23,1%	100,0%
	1 ženské	Count	2	11	13
		% within pohlavi	15,4%	84,6%	100,0%
Total		Count	12	14	26
		% within pohlavi	46,2%	53,8%	100,0%

Nejprve využijme informaci o náušnici

	naušnice		Total
	nenosí	nosí	
mužské	10	3	13
ženské	2	11	13
	12	14	26

- Šance, že osoba nosící náušnici je žena =
 $O_{(\text{žena}|\text{nosí})} = 11/3 = 3,7:1 \dots\dots\dots P_{(\text{žena}|\text{nosí})} = 0,79$
- Šance, že osoba nenosící náušnici je žena =
 $O_{(\text{žena}|\text{nenosí})} = 2/10 = 0,2:1 \dots\dots\dots P_{(\text{žena}|\text{nenosí})} = 0,17$
- Nosí-li náušnici, je asi 18krát větší šance, že je to žena, než když ji nenosí.

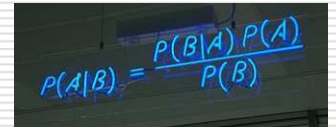
■ **Poměr šancí** za dvou podmínek (**odds ratio, OR**) =

$$OR = \frac{O_{(\text{žena}|\text{nosí})}}{O_{(\text{žena}|\text{nenosí})}} = \frac{3,7}{0,2} \cong 18,3$$

A co informace o emočních adjektivech?

emoce			
pohlaví	M	N	SD
mužské	4,46	13	2,634
ženské	9,00	13	3,291
Total	6,73	26	3,726

- Z těch, kdo mají $EA=8$, je $7/8$ žen a $1/8$ mužů $O(\text{žena}|e=8)=7$...ale dat je málo a nevyužíváme informaci o rozložení
- Předpokládáme-li v populaci normální rozložení...
 - $P(EA \geq 8|\text{žena}) = 1 - \text{norm.dist}(8; 9; 3,291) = 0,62$
 - $P(\text{ž}|EA \geq 8) = [P(EA \geq 8|\text{ž}) * P(\text{ž})] / [P(EA \geq 8|\text{ž}) * P(\text{ž}) + P(EA \geq 8|m) * P(m)] =$
 $= [0,62 * 0,5] / [0,62 * 0,5 + 0,09 * 0,5] = 0,87$... $O(\text{ž}|EA \geq 8) = 6,9$
 - pro $EA \geq 9$ je $O(\text{ž}|EA \geq 9) = 11,8$
 - $OR(e \geq 9 \text{ ku } e \geq 8) = 11,8 / 6,9 = 1,7$
 - **Poměr šancí** spojený s nárůstem **e.a.** o 1 je 1,7


$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Uff, a to jsme nevzali v potaz možnou souvislost mezi nošením náušnic a emočními adjektivy....

Logistická regrese

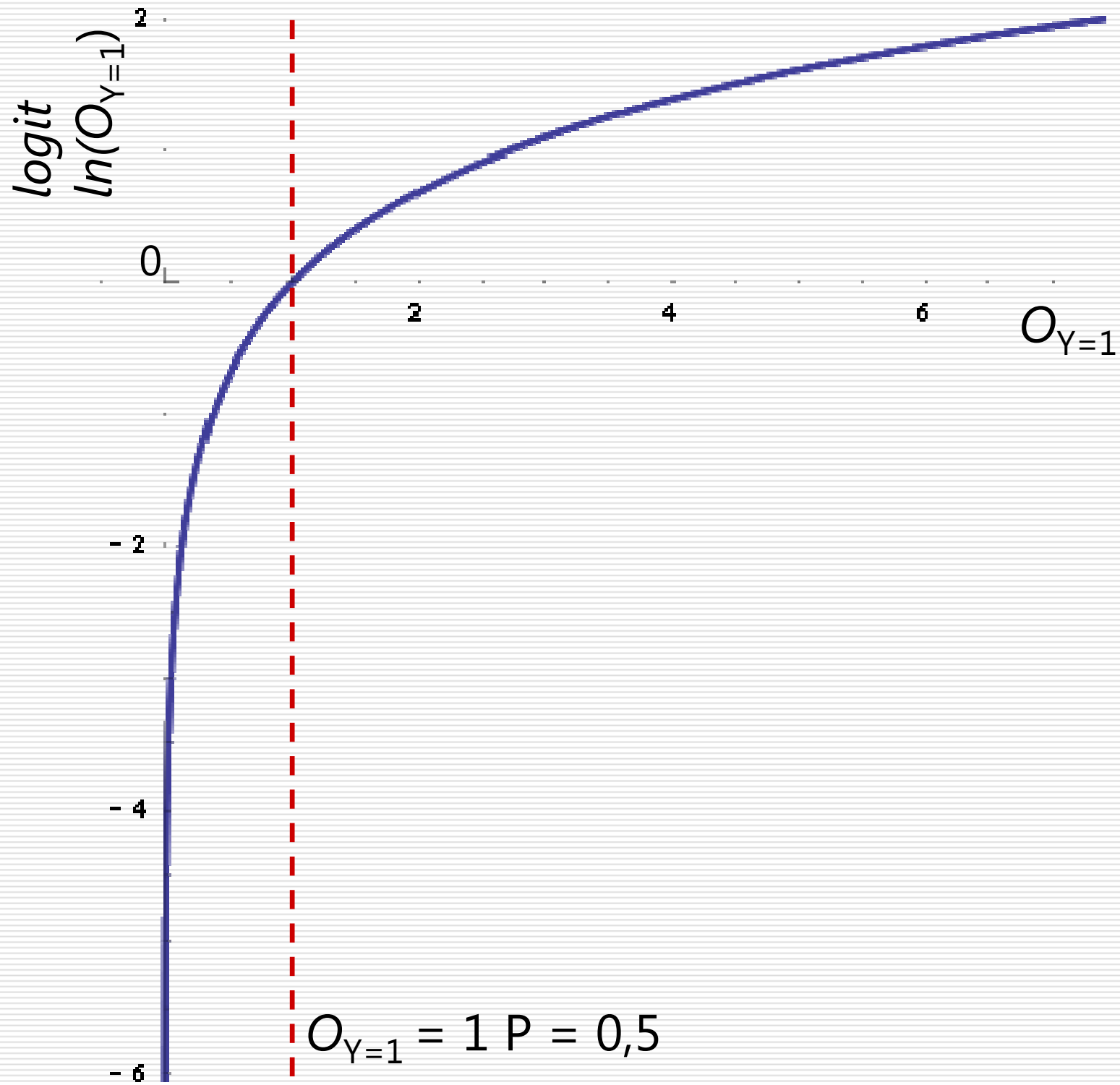
- Rozšíření lineární regrese na dichotomické závislé
 - není to lineární regrese, protože nejde o lineární vztah
 - Závislou kódujeme 1 (jev nastal) a 0 (jev nenastal)
 - *Ideově* je závislou proměnnou **pravděpodobnost toho, že jev nastal(nastane)**
 - Pomocí prediktorů predikujeme, jaká je pravděpodobnost, že jev nastane.
-

Technický základ logistické regrese 1

- šance $O_{Y=1} = P_{Y=1}/P_{Y \neq 1} = P_{Y=1}/(1-P_{Y=1})$
- $\ln O_{Y=1}$ se jmenuje **logit** ($P_{Y=1}$)

Malé matematické opáčko

- \ln je přirozený logaritmus, tj. log se základem $e=2,7\dots$
 - $\ln(x) = \log_e(x) = y$ - číslo, kterým když umocníme e , dostaneme x
... $e^y=x$
 - Protože cokoli umocněné na nultou = 1, $\ln(1) = 0$
 - Pro $x < 1$, $\ln(x) < 0$; pro $x > 1$, $\ln(x) > 0$
-



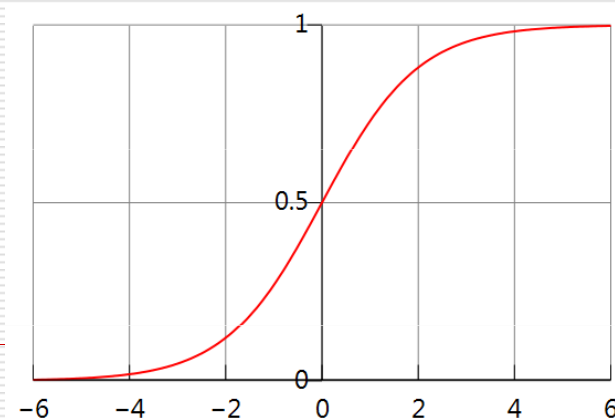
Proč tak složitě?

Závislá jako pravděpodobnost má měřítko v rozsahu $\langle 0;1 \rangle$. Kombinace prediktorů má ale neomezený rozsah $(-\infty;\infty)$.

Proto změníme měřítko závislé

1. Místo P použijeme O s měřítkem $\langle 0; \infty \rangle$
2. Pomocí logaritmu změníme měřítko na $(-\infty;\infty)$.

Také lze říci, že jde o linearizaci vztahu.



Technický základ logistické regrese 1

- šance $\mathbf{O}_{Y=1} = P_{Y=1}/P_{Y \neq 1} = P_{Y=1}/(1-P_{Y=1})$
- $\ln \mathbf{O}_{Y=1}$ se jmenuje **logit** ($P_{Y=1}$)
- 2(3) ekvivalentní rovnice modelu logistické regrese

$$\ln \mathbf{O}_{Y=1} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_m X_m$$

$$P_{Y=1} = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \dots + b_m X_m)}} = \frac{e^{(b_0 + b_1 X_1 + \dots + b_m X_m)}}{1 + e^{(b_0 + b_1 X_1 + \dots + b_m X_m)}}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	nausnice	2,909	1,012	8,260	1	,004	18,333
	Constant	-1,609	,775	4,317	1	,038	,200

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

$$P_{Y=\text{žena}} = \frac{1}{1 + e^{-(-1,6 + 2,9NA)}}$$

$$\ln O_{Y=\text{žena}} = -1,6 + 2,9n\text{áúšnice}$$

- Pro náušnice=1 ... $P_{(\text{žena}|\text{náušnice})} = 0,79$ $O = 3,7$
- Kdyby neměl náušnici ... $P = 0,17$ $O = 0,2$
- Změna náušnice z 1 na 0 způsobila
18násobný pokles šancí ... $\exp(B) \dots e^b$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	emoce	,486	,185	6,875	1	,009	1,625
	Constant	-3,219	1,305	6,088	1	,014	,040

$$P_{Y=\text{žena}} = \frac{1}{1 + e^{-(-3,2 + 0,5EM)}}$$

$$\ln O_{Y=\text{žena}} = -3,2 + 0,5\text{emoce}$$

- Pro emoce=8 ... $P_{(\text{žena}|e=8)} = 0,66$ $O = 1,9$
- Pro emoce=9 ... $P = 0,76$ $O = 3,2$
- Změna emocí z 8 na 9 způsobila 1,6násobný nárůst šancí ... stejně jako jakékoli změna o 1

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	nausnice	2,147	1,144	3,523	1	,061	8,556	,909	80,501
	emoce	,387	,199	3,791	1	,052	1,472	,997	2,173
	Constant	-3,797	1,545	6,041	1	,014	,022		

$$P_{Y=\text{žena}} = \frac{1}{1 + e^{-(-3,8 + 0,4EM + 2,1NA)}}$$

$$\ln O_{Y=\text{žena}} = -3,80 + 0,39\text{emoce} + 2,15\text{náušnice}$$

- Pro náušnice=1 a emoce=8 ... $P=0,81$ $O=4,2$
- Kdyby neměl náušnici ... $P=0,33$ $O=0,50$
- Změna náušnice z 1 na 0 (bez změny e.a.)
způsobila 8,5násobný pokles šancí ... e^b

Technický základ logistické regrese 2

Jak spočítáme regresní váhy, které vyústí v nejlepší predikci pravděpodobnosti $Y = 1$?

- nespočítáme, odhadneme (zapomeňme na nejmenší čtverce)
 - odhad metodou **maximální věrohodnosti** (maximum-likelihood estimation)
 - Algoritmus, který postupným přibližováním...
 - ...dochází k takovým váhám, s nimiž je podmíněná pravděpodobnost získání dat, která jsme získali, nejvyšší možná : $P(\text{data} | b_0, b_1, \dots, b_m) = \max$
 - likelihood = jiné slovo pro podmíněnou p-nost
-

Residua v logistické regresi

- ❑ **Rezidua** jsou rozdíl mezi predikovanou P a skutečnou hodnotou Y (1 nebo 0)
 - ❑ Nejsou normálně rozložena
 - ❑ Normálně nejsou rozložena ani standardizovaná rezidua
 - ❑ Pro detekci *outlierů* jsou tak vhodná **studentizovaná rezidua**, která mají přibližně normální rozložení, a s ním spojená očekávání.
-

Jak dobře regrese predikuje?

For dummies (matematically to kulhá)

- Základem je opět rozdíl mezi skutečnou hodnotou Y a odhadem $P_{Y=1}$
 - Je-li $Y=1$, pak je rozdíl $1-P$
 - Je-li $Y=0$, pak je rozdíl $0-P$

Jeden z rozdílů je vždy záporný bez ohledu na to, jak odečítáme.

Namísto umocňování zde rozdíly logaritmujeme a sečteme.

Získáme velké číslo, které funguje jako suma čtverců reziduí v lineární regresi.

$$LL = \sum^N [Y_i \ln P_{Y=1} + (1 - Y_i) \ln(1 - P_{Y=1})]$$

Jak dobře regrese predikuje?

- Měřítkem zdařilosti regrese je logaritmus p-nosti dat při odhadnutých hodnotách regresních koeficientů:

log-likelihood

$$LL = \sum_{i=1}^N [Y_i \ln P_{Y=1} + (1 - Y_i) \ln(1 - P_{Y=1})]$$

- **LL** sumíruje shodu mezi odhadem a daty
 - maximem je 0, minimem je $-\infty$
 - častěji se udává jako **-2LL**, tj. vynásobený -2
 - **-2LL** se říká **deviance** (0 až ∞)
 - má chíkvadrát rozložení
 - **reportujeme Model chi-square, df, p**
-

Statistické testy 1

Predikuje regrese lépe než *nic*?

- nic = základní model (baseline model) = predikujeme všem 0 nebo 1, podle toho, co z toho se vyskytuje častěji = $P_{Y=1}$ je pro všechny lidi stejná
- Potom můžeme srovnat model s prediktory s tímto základním modelem.
 - rozdíl $-2LL$ obou modelů má χ^2 rozložení s df =počet prediktorů

$$\chi^2 = -2LL_{\text{náš model}} - (-2LL_{\text{základní model}})$$

$$df = m_{\text{náš model}} - m_{\text{základní model}}$$

- tj. je-li $1 - \text{CHISQ.DIST}(\chi^2; df) < 0,05$, predikuje model lépe než nic
 - Podobně můžeme srovnávat i modely s různým počtem prediktorů mezi sebou
-

Nedalo by se to trochu zjednoduřit?

-2LL lze převést na ukazatele podobné R^2

$$LL=0 \implies R^2=1 \dots\dots LL=-\infty \implies R^2=0$$

- R_L^2 Hosmera a Lemeshowa
- R_{CS}^2 Coxe a Snella ($\max R_{CS}^2 < 1$)
- R_N^2 Nagelkerkeho ($R_{CS}^2 / \max R_{CS}^2$)

Nabývají hodnot od 0 do 1.

Udávají jak moc díky prediktorům klesl -2LL

Není to úplně totéž, co R^2 v lineární regresí!

Jak dobře regrese predikuje? Klasifikační tabulka.

Classification Table^{a,b}

Observed		Predicted		
		pohlavi		Percentage Correct
		0 mužské	1 ženské	
Step 0	pohlavi 0 mužské	0	13	0,0
	1 ženské	0	13	100,0
Overall Percentage				50,0

Classification Table^a

Observed		Predicted		
		pohlavi		Percentage Correct
		0 mužské	1 ženské	
Step 1	pohlavi 0 mužské	11	2	84,6
	1 ženské	2	11	84,6
Overall Percentage				84,6

a. The cut value is 0,500

Interpretace regresních koeficientů

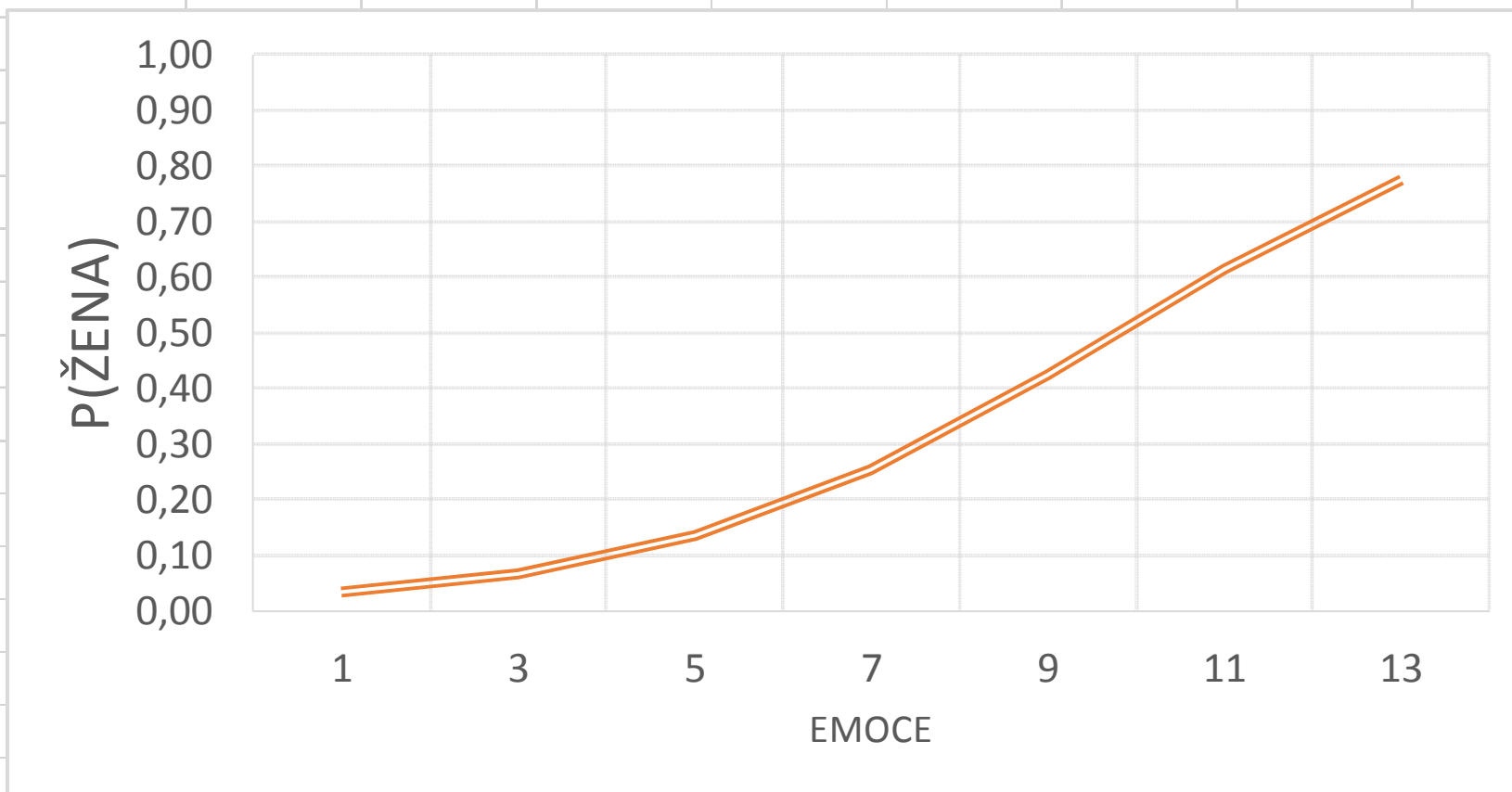
- U kategorických prediktorů (indikátorově kódovaných) udává ***expB*** poměr šancí pro indikovanou hodnotu vs. referenční hodnotu.
 - U spojitých prediktorů udává ***expB*** poměr šancí (nárůst) spojený s jednotkovým rozdílem na škále prediktoru.
 - Standardní velikost účinku vyjádřená OR je někdy zrádná (neznáme základ jako u procent)
 - Proto počítáme **rozdíl p-ností predikovaných** pro dvě různé (typické) hodnoty určitého prediktoru.
-

Pro interpretaci je dobré vykreslit vztah mezi Y a X do grafu...

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	emoce	0,387	0,199	3,791	1	0,052	1,472
	nausnice	2,147	1,144	3,523	1	0,061	8,556
	Constant	-3,797	1,545	6,041	1	0,014	0,022

a. Variable(s) entered on step 1: emoce, nausnice.

	emoce =	1	3	5	7	9	11	13
náušnice =	0	0,03	0,07	0,13	0,25	0,42	0,61	0,77



Statistické testy 2

Testy jednotlivých prediktorů

- Waldův test: $z = b / SE(b)$
 - SPSS: Wald = z^2 , Wald $\sim \chi^2(df)$
 - při velkých b nadhodnocuje SE
 - i tak je dobré **uvádět 95% CI pro expB**
 - Robustnější alternativou je χ^2 test zhoršení modelu po vyřazení daného prediktoru (tzv. **likelihood-ratio test, LRT**)
-

Další indikátory kvality modelu

- Klasifikační tabulka
 - srovnání predikovaného a skutečného stavu
 - „reality-check“, i krásně signifikantní model může neuspokojivě predikovat
 - Hosmer-Lemeshow Goodness of Fit Test
 - také srovnává predikované a pozorované hodnoty závislé
 - GoF test >> nechceme, aby byl signifikantní
 - Klasifikační diagram (classification plot)
 - Diagnostika reziduí a vlivných případů (jako v LinReg)
-

Praktické problémy

- Regresní koeficienty se nevypočítávají, ale iteračně odhadují.
 - Iterace nemusí vždy proběhnout úspěšně
 - nemusí konvergovat
 - mohou se vyskytnout bláznivé hodnoty
 - Problematické výsledky naznačují nedostatky v datech
 - při absenci některé z kombinace hodnot prediktorů a závislé
 - při dokonalé predikci
 - LR je náročná na velikost vzorku
-

Předpoklady logistického modelu

- ❑ Není jich mnoho
 - ❑ Linearita – předpoklad lineárního vztahu mezi spojitými prediktory a logitem závislé.
 - ❑ Nezávislost reziduí
 - ❑ Implicitně dostatek dat – měly by se vyskytovat všechny kombinace kategorických prediktorů
 - ❑ Multikolinearita je stejným problémem jako u LinReg
-

Obecně budování modelu

- Vzhledem k nárokům na velikost vzorku větší tlak na jednoduchost modelu
 - *Explorace*: Vložit všechny prediktory a postupně ubírat – cílem je parsimonie (úspornost)
 - *Testování hypotéz*: vložit, co implikuje teorie, smysluplně po blocích
-

Reportování

Field 19.7

Kam dál?

- ordinální regrese
 - multinomiální/multinomická regrese

 - Generalizovaný lineární model
-

Shrnutí

- ❑ Logistická regrese umožňuje predikovat dichotomii – zda se něco stane, či ne
 - ❑ Na straně prediktorů se oproti lineární regresi nic nemění – je toho dosaženo transformací závislé proměnné na logity
 - ❑ Transformace komplikuje interpretaci regresních koeficientů
 - ❑ Ubylo předpokladů
-

Seminární úkol

- Data *Erasmus*
 - Predikujeme, zda během Erasmovského pobytu dojde k rozchodu
 - Prediktory jsou
 - délka vztahu
 - spokojenost ve vztahu
 - pohlaví
 - attachmentový styl
-