PSY252

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 83% 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 – $P(\text{žena}|\text{nosí})=?$
- $P(\text{ž}|\text{n}) = P(\text{n}|\text{ž})P(\text{ž})/P(\text{n}) =$

  $= P(\text{n}|\text{ž})P(\text{ž})/(P(\text{n}|\text{ž})P(\text{ž})+P(\text{n}|\text{m})P(\text{m}))=$

  $=0{,}85*\mathbf{0{,}5}/(0{,}85*0{,}5+0{,}23*0{,}5) = \mathbf{0{,}79}$

```
CROSSTABS
  /TABLES=pohlavi BY nausnice
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

pohlavi * nausnice Crosstabulation

			nausnice		Total
			0 nenosí	1 nosí	
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

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

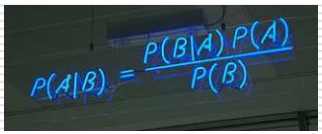
- Šance, že osoba nosící náušnici je žena=
  $O_{(\text{žena}|\text{nosí})} = 11/3 = 3{,}7{:}1$ ..... $P_{(\text{žena}|\text{nosí})} = 0{,}79$
- Šance, že osoba nenosící náušnici je žena=
  $O_{(\text{žena}|\text{nenosí})} = 2/10 = 0{,}2{:}1$ ..... $P_{(\text{žena}|\text{nenosí})} = 0{,}17$
- Nosí-li 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			
pohlavi	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 e=8, je 7/8žen 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…
  - $P(e\geq 8|\text{žena})=\text{normsdist}(-0,3)=0,62$
  - $P(\text{ž}|e\geq 8)=[P(e\geq 8|\text{ž})*P(\text{ž})]/[P(e\geq 8|\text{ž})*P(\text{ž})+P(e\geq 8|\text{m})*P(\text{m})]=$
    $=[0,62*0,5]/[0,62*0,5+0,09*0,5]=0,87$ … $O(\text{ž}|e\geq 8)=6,9$
  - pro e≥9 je $O(\text{ž}|e\geq 9)=11,8$
  - OR(e≥9 ku e≥8)=11,8/6,9=1,7
  - **Poměr šancí** spojený s nárůstem ***e.a.*** o 1 je 1,7



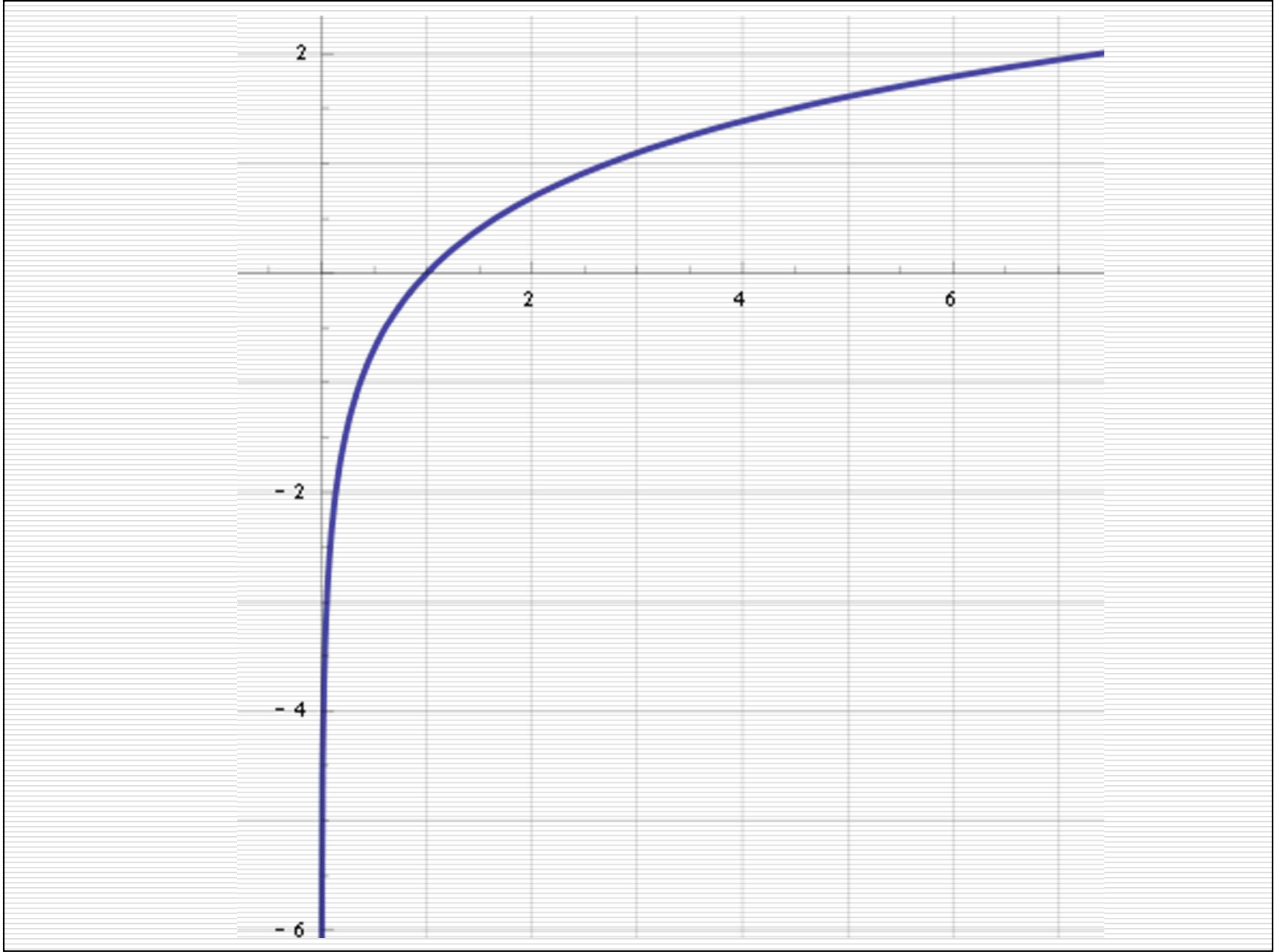
Uff, a to jsme nevzali v potaz možnou souvislost mezi nošením náušnic a emočními adjektivy….

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


2
2
4
6
– 2
– 4
– 6

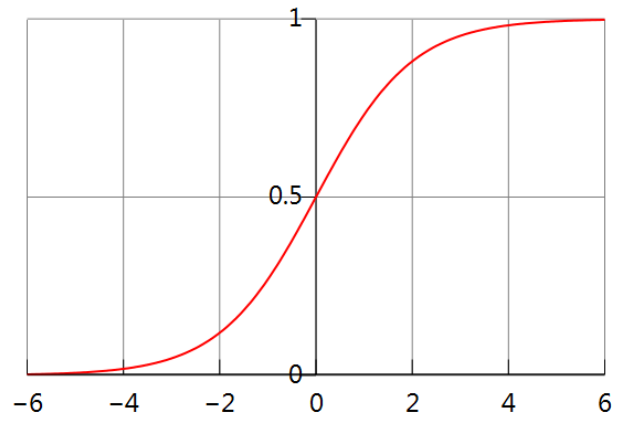
Proč tak složitě?

Závislá jako pravděpodobnost má měřítko v rozsahu <0;1>. Kombinace prediktorů má ale rozsah (−∞;∞).

Proto změníme měřítko závislé

1. Místo P použijeme O s měřítkem <0; ∞)
2. Pomocí logaritmu změníme měřítko na (−∞;∞).

Také lze říci, že jde o linearizaci vztahu.



Technický základ logistické regrese 1

- **šance $O_{Y=1}$** = $P_{Y=1}/P_{Y\neq1} = P_{Y=1}/(1-P_{Y=1})$
- **ln $O_{Y=1}$** se jmenuje **logit ($P_{Y=1}$)**
- Ekvivalentní rovnice modelu logistické regrese

ln $O_{Y=1}$ = $b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$

$O_{Y=1} = e^{(b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m)}$

$$P_{Y=1} = \frac{1}{1+e^{-(b_0+b_1X_1+\ldots+b_mX_m)}} = \frac{e^{(b_0+b_1X_1+\ldots+b_mX_m)}}{1+e^{(b_0+b_1X_1+\ldots+b_mX_m)}}$$

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.

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

$$\ln O_{Y=\check{z}ena} = -1,6 + 2,9\text{náušnice}$$

- Pro náušnice=1 … $P_{(\check{z}ena|n\acute{a}u\check{s}nice)}=0,79$ $O=3,7$
- Kdyby neměl náušnici … $P=0,17$ $O=0,2$
- Změna náušnice z 1 na 0 způsobila 18násobný pokles šancí …. exp(B)… 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

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

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

- Pro emoce=8 … $P_{(\check{z}ena|e=8)}=0{,}66$ $O=1{,}9$
- 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		

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

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

- 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)
  - Výpočetně složitý algoritmus
  - Dochází k takovým regr. koef., s nimiž je podmíněná pravděpodobnost získání dat, která jsme získali, nejvyšší možná : P (data$|b_0,b_1,..,b_m$) = max
  - likelihood = podmíněná p-nost P(D|H) pro různé H

Jak dobře regrese predikuje?

- Likelihood je měřítkem zdařilosti regrese v logaritmované podobě: **log-likelihood**

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

- ***LL*** sumíruje shodu mezi odhadem a daty
  - maximem je 0, minimem je -∞
  - 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 – **likelihood ratio test, LRT**.
  - 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-CHISQ.DIST(χ^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 zjednodušit?

-2LL lze převést na ukazatele podobné R^2

$-2LL=0$ … $R^2=1$ a $-2LL=\infty$ … $R^2=0$

- 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 regresi!

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.

Statistické testy 2
Testy jednotlivých prediktorů

- Waldův test: $z = b/\mathrm{SE}(b)$
  - SPSS: Wald=z^2, Wald~$\chi^2(df)$
  - 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**)

Další indikátory kvality modelu

- Klasifikační tabulka – úspěšnost predikce
  - srovnání predikovaného a skutečného stavu
  - „reality-check", i krásně signifikantní model může neuspokojivě predikovat
- Hosmer-Lemeshow Goodness of Fit Test
  - také srovnává predikovné 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 regrese

- Generalizovaný lineární model

Seminární úkol

- Connie data
- Predikujeme b05h (dobrovolničení)
  - vzděláním otce
  - hodnotami: hod_mat hod_eco hod_infl sko_zap hod_edu
  - *?* je efekt hod_eco moderován generací (1995-2010)?
- Popsat výsledný model
  - Kvalita modelu – testy, klasifikační úspěšnost, předpoklady, vlivné případy
  - Vliv prediktorů – testy, interpretace, ilustrovat predikovanými pravděpodobnostmi