Kateřina Škařupová
43275
skarupov@fss.muni.cz

Přemysl Maršík
13477
13477@mail.muni.cz

Analýza kategorizovaných dat v sociologii
Úkol č. 2

Zadali jsme Goodmanova data do staty ve formě uvedené v zadání úkolu, tedy:

Rasa: Afroameričané=0, bílí Američané=1
Origin: sever=0, jih=1
Lokace: severní=0, jižní=1 (jednotka)
Preference: severní=0, jižní=1 (jednotka)

Pomocí logitové regrese jsme vytvořili lineární model popisující preference respondentů.
Odhad nejúspornějšího modelu:

preference	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
_cons	-.0164265	.0223113	-0.74	0.462	-.0601558	.0273029

Tento model nebere v potaz vysvětlující proměnné, obsahuje pouze pravděpodobnost preference umístění. Konstanta odpovídá pravděpodobnosti preference jižní jednotky 49,59 % mezi všemi vojáky.

Další model obsahuje všechny vysvětlující proměnné:

preference	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
rasa	.7616412	.0619217	12.30	0.000	.6402768	.8830055
origin	2.613202	.0623686	41.90	0.000	2.490961	2.735442
lokace	1.548657	.0627151	24.69	0.000	1.425738	1.671576
_cons	-2.75682	.0785685	-35.09	0.000	-2.910812	-2.602829

Modely s interakcí:
Vytvořili jsme 12 modelů s interakcemi. Uvážili jsme totiž, že zavedením jednoduché interakce mezi proměnnými, např.:

Rasa*origin=nováproměnná,

bychom potlačili možné interakce mezi jednotlivými variantami proměnných. To znamená, že pro každou dvojici původních proměnných vytvoříme všechny možné kombinace, a ke každé z nich vytvoříme interakční proměnnou, např.:

Rasa a původ:

Bílý Američan, původem z jihu
NewVar1=rasa*origin
Afroameričan, původem z jihu
NewVar2=(1-rasa)*origin
Bílý Američan, původem ze severu

NewVar3=rasa*(1-origin)
Afroameričan, původem ze severu
NewVar4=(1-rasa)*(1-origin)

Použili jsme zjednodušené modely, ve kterých vystupovala pouze interakce a třetí proměnná, nikoliv však samotné interagující proměnné. Tyto modely jsme porovnali ve čtveřicích mezi sebou a vybrali nejvhodnější. Předpokládali jsme tedy, že je-li zjednodušený model(symbolicky):
logit depvar varA varB*varC věrohodnější než model(např.):
logit depvar varA varNotB*varC,
pak bude i doplněný model:
logit depvar varA varB varC varB*varC věrohodnější, než
logit depvar varA varB varC varNotB*varC.

Nutno podotknout, že korektní důkaz jsme neprovedli, uvažujeme pouze, že je-li určitá interakce lepším modelem dat, než jiná ve zjednodušeném případě, pak doplněný model sednutí (fit) sice zlepší v obou případech, nezamíchá však s pořadím.

Tímto způsobem jsme nalezli v každé ze tří čtveřic nejdůležitější binární interakci mezi proměnnými, výslovně tedy mezi

1) **původem z jihu a umístěním v jižní jednotce**
2) **Afroameričan a původem ze severu**
3) **Afroameričan a umístěný na severu**

Tyto tři modely, dále už doplněné i o samotné interagující proměnné, jsme porovnali mezi sebou a s modelem bez interakce. Nejlépe vystihoval data model s interakcí mezi **původem z jihu a umístěním v jižní jednotce.** Tedy symbolicky **logit preference rasa origin lokace origin*lokace.**

preference	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
rasa	.7426203	.0622675	11.93	0.000	.6205782	.8646624
origin	2.180223	.1065738	20.46	0.000	1.971342	2.389103
lokace	1.219707	.0891185	13.69	0.000	1.045038	1.394376
sou_sou~n	.5957729	.1216427	4.90	0.000	.3573576	.8341882
_cons	-2.490725	.0921376	-27.03	0.000	-2.671312	-2.310139

Míra fitu modelu na data:

Log-Lik Intercept Only:	-5569.860	Log-Lik Full Model:	-4014.847
D(8031):	8029.694	LR(4):	3110.025
Prob > LR:	0.000		
McFadden's R2:	0.279	McFadden's Adj R2:	0.278
Maximum Likelihood R2:	0.321	Cragg & Uhler's R2:	0.428
McKelvey and Zavoina's R2:	0.404	Efron's R2:	0.350
Variance of y*:	5.524	Variance of error:	3.290
Count R2:	0.776	Adj Count R2:	0.547
AIC:	1.000	AIC*n:	8039.694
BIC:	-64182.542	BIC':	-3074.059

Tentýž model v odds ratio:

preference	Odds Ratio	Std. Err.	z	P>z	[95% Conf.	Interval]
rasa	2.101435	.1308511	11.93	0.000	1.860003	2.374204
origin	8.848276	.9429942	20.46	0.000	7.180305	10.90371
lokace	3.386195	.3017726	13.69	0.000	2.843506	4.032457
sou_sou~n	1.814433	.2207125	4.90	0.000	1.429547	2.302944

Interpretace modelu s interakcí není zcela triviální, protože model již není lineární, má totiž tvar

$$y = ax_1 + bx_2 + cx_3 + dx_2x_3 + e$$

Uvážíme-li však tuto nelinearitu, a započítáme-li vliv interceptu(zde jako faktor **f=0,083=exp(_cons)**)z koeficientů pro odds ratio můžeme určit, že:

- 2,1krát více bílých Američanů než Afroameričanů volí jižní camp. Tato proměnná neinteraguje s ostatními.
- U vojáka, u nějž známe pouze původ (řekněme jižní), můžeme s 8,8krát větší šancí očekávat preferenci jižní jednotky, než u vojáka ze severu.
- U vojáka, u nějž známe pouze umístění (řekněme jižní), můžeme s 3,4krát větší šancí očekávat preferenci jižní jednotky než u vojáka ze severní jednotky.
- U Afroameričana původem z jihu, umístěného na severu bude šance, že preferuje jižní jednotku 0,73, protože 0,73=8,8*f. U bělocha původem z jihu, umístěného na severu bude šance preference jižní jednotky 1,53 (=2,1*8,8*f), pravděpodobnost tedy 60,5 %.
- Avšak pro vojáka z jihu, umístěného na jihu vstupuje do hry interakce a šance pro preferenci jižní jednotky je u Afroameričanů 4,47 (=8,8*3,4*1,8*f), pravděpodobnost tedy 81,7 %. Pro bílé Američany pak šance činí 9,39 a pravděpodobnost 90,4 %.
- Atd..

Predikujeme-li příslušné hodnoty ve STATě, naše závěry se potvrzují. Nesmíme samozřejmě zapomenout, že interakční proměnná není nezávislá:

predikce :Afroameričan,ze severu,na severu

```
Pr(y=1x):          0.0765   95% ci: (0.0647,0.0903)
Pr(y=0x):          0.9235   95% ci: (0.9097,0.9353)
              rasa        origin        lokace  south_sout~n
x=            0             0             0             0
```

predikce :bílý Američan,ze severu,na severu

```
Pr(y=1x):          0.1483   95% ci: (0.1302,0.1684)
Pr(y=0x):          0.8517   95% ci: (0.8316,0.8698)
              rasa        origin        lokace  south_sout~n
x=            1             0             0             0
```

predikce :bílý Američan,z jihu ,na severu

```
Pr(y=1x):          0.6064   95% ci: (0.5686,0.6430)   viz interpretace
Pr(y=0x):          0.3936   95% ci: (0.3570,0.4314)
               rasa        origin        lokace  south_sout~n
x=             1             1             0             0
```

predikce :bílý Američan,ze severu,na jihu

```
Pr(y=1x):          0.3709   95% ci: (0.3484,0.3939)
Pr(y=0x):          0.6291   95% ci: (0.6061,0.6516)
               rasa        origin        lokace  south_sout~n
x=             1             0             1             0
```

predikce :bílý Američan,z jihu,na jihu

```
Pr(y=1x):          0.9044   95% ci: (0.8918,0.9157)   viz interpretace
Pr(y=0x):          0.0956   95% ci: (0.0843,0.1082)
               rasa        origin        lokace  south_sout~n
x=             1             1             1             1
```

predikce :Afroameričan,z jihu,na jihu

```
Pr(y=1x):          0.8183   95% ci: (0.8027,0.8330)   viz interpretace
Pr(y=0x):          0.1817   95% ci: (0.1670,0.1973)
               rasa        origin        lokace  south_sout~n
x=             0             1             1             1
```