vaja III

LOGISTIČNA REGRESIJA

Vsi podatki, potrebni za to vajo, se nahajajo v mapi ..\Statistika\SSMM\Vaja3

PRIMER 1: PREDSEDNIŠKE VOLITVE V ZDA LETA 1992

Raziskovalno vprašanje: Napovedovanje glasov za W.J.Clintona na podlagi starosti in spola in izobrazbe.
Datoteka: Voter.sav

1. Oglejte si glasovanje za Clintona glede na spol in izračunajte razmerje obetov za ženske glede na moške ter "relativno tveganje" glasovanja za Clintona za ženske glede na moške. (Uporabljajte spremenljivko Clinton, kjer smo združili v eno kategorijo volilci, ki so volili za Busha ali za Perrota, kodiranje: 1 za Clintona, 0 za ostale.)

(v SPSS-u : Analzye -> Descriptive Statistics -> Cross Tabulations -> … in Format… (ascending order of rows))

Vote for Clinton * RESPONDENTS SEX Crosstabulation

Count

		RESPONDENTS SEX		Total
		Female	male	
Vote for Clinton	Clinton	a	b	a+b
	Bush or Perot	c	d	c+d
Total		a+c	b+d	

Verjetnosti p: P(Clinton)=
p_M = P(Clinton | Spol=Moski)=
p_F =P(Clinton | Spol= Zenski)=

Relativno tveganje glasovanja za Clintona za ženske glede na moške (relative risk, RR):

$$RR=\frac{p_F}{p_M}$$

95% IZ za RR je : 1,16 do 1,36
Dobili smo ga z naslednjem postopkom

log(RR) je normalno porazdeljen in $SE(\log(RR)) = \sqrt{\frac{1-\frac{a}{a+b}}{a} + \frac{1-\frac{c}{c+d}}{c}}$

95% IZ za log(RR) : $\log(RR) - 1.96 * SE(\log(RR)), \log(RR) + 1.96 * SE(\log(RR))$

95% IZ za RR: $e^{\log(RR)-1.96*SE(\log(RR))}, e^{\log(RR)+1.96*SE(\log(RR))}$

Obet glasovanja za Clintona za ženske (odds) : $\frac{p_F}{1-p_F} =$

Obet glasovanja za Clintona za moške (odds) : $\frac{p_M}{1-p_M} =$

Razmerje obetov glasovanja za Cintona za ženske glede na moške (odds ratio, OR):

$$OR = \frac{\frac{p_F}{1-p_F}}{\frac{p_M}{1-p_M}} = \frac{ad}{bc} =$$

95% IZ za OR je: 1,39 do 2,02

Dobili smo ga z naslednjem postopkom :

log(OR) je normalno porazdeljen in $SE(\log(OR)) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$ =0,095, log(OR)=0,517

95% IZ za log(OR) $\log(OR) - 1.96 * SE(\log(OR)), \log(OR) + 1.96 * SE(\log(OR))$

95% interval zaupanja za OR $e^{\log(OR)-1.96*SE(\log(OR))}, e^{\log(OR)+1.96*SE(\log(OR))}$

Razmerje obetov in relativno tveganje

$RR = \frac{p_F}{p_M} = \frac{\frac{p_F(1-p_F)}{1-p_F}}{\frac{p_M(1-p_M)}{1-p_M}} = OR\frac{1-p_F}{1-p_M} \cong OR$, če sta p_F in p_M sta majhni.

2. Z modelom logistične regresije si oglejte vpliv spola na glasovanje za Clintona. Podrobno interpretirajte rezultate.

Coding of Sex: 1=M, 2=F
Dependent Variable Encoding: 1=Cllinton, 0=Bush or Perot

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	sex	,517	,095	29,730	1	,000	1,676	1,392	2,019
	Constant	-,843	,156	29,232	1	,000	,430		

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

Kako lahko interpretiramo oceno za regresijski koeficient za spol? Ali obstaja kakšna povezanost z rezultati, ki smo jih dobili pri prejšnji nalogi?
(Določite razmerje obetov za žensko glede na moškega in 95% interval zaupanja.)

Kaj je ničelna domneva?

Določite razmerje obetov za moškega glede na žensko in 95% interval zaupanja.

3. Z modelom logistične regresije si oglejte povezanosti **starosti in spola** z glasovanjem za Clintona. (SPSS: Analyze-> Regression -> Binary Logistic).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sex	,516	,095	29,447	1	,000	1,675
	age	,011	,003	13,335	1	,000	1,011
	Constant	-1,349	,210	41,251	1	,000	,259

a. Variable(s) entered on step 1: sex, age.

Ali lahko interpretiramo regresijski koeficient za spol na isti način, kot smo ga interpretirali v prejšnjem modelu?

Izračunajte razmerje obetov za dva moška, katerih starost se razlikuje za 5 let.

Izračunajte razmerje obetov za dve ženski, katerih starost se razlikuje za 5 let.

Izračunajte pričakovano verjetnost, da bo ženska stara 40 let volila za Clintona.

4. Z modelom logistične regresije si oglejte povezanost **števila let šolanja (spremenljivka educ)** z glasovanjem za Clintona in jih primerjajte z rezultati, ki jih dobimo, ko uporabljamo **stopnjo izobrazbe** (spremenljivka **degree)**. Primerjajte rezultate modela z opazovanimi rezultati.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	educ	-,022	,016	1,805	1	,179	,978
	Constant	,273	,233	1,374	1	,241	1,314

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

Katero predpostavko naredimo glede povezanosti med leti šolanja in verjetnostjo glasovanja za Clintona?

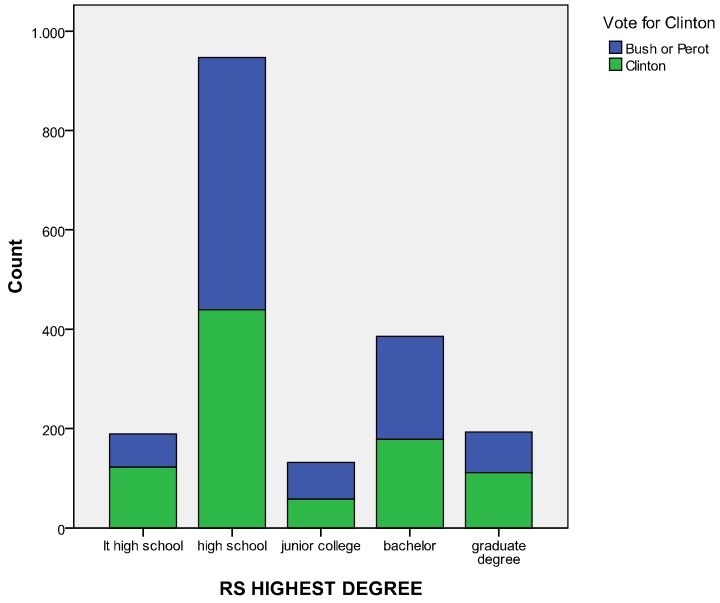
Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1[a]	degree			28,495	4	,000	
	degree(1)	,297	,211	1,984	1	,159	1,345
	degree(2)	-,449	,160	7,914	1	,005	,638
	degree(3)	-,546	,228	5,747	1	,017	,579
	degree(4)	-,459	,178	6,648	1	,010	,632
	Constant	,303	,146	4,324	1	,038	1,354

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

- Kodiranje: 0 = "less than high school" ; 1 = "high school" ; 2 = "junior college"; 3 = "bachelor" ; 4 = "graduate degree"

Opazovani rezultati



Vote for Clinton * RS HIGHEST DEGREE Crosstabulation

			RS HIGHEST DEGREE					Total
			lt high school	high school	junior college	bachelor	graduate degree	
Vote for Clinton	Clinton	Count	122	439	58	178	111	908
		% within RS HIGHEST DEGREE	64,6%	46,4%	43,9%	46,1%	57,5%	49,2%
	Bush or Perot	Count	67	508	74	208	82	939
		% within RS HIGHEST DEGREE	35,4%	53,6%	56,1%	53,9%	42,5%	50,8%
Total		Count	189	947	132	386	193	1847

4. Dodatna naloga
   V model vstavite število let šolanja ter kvadrat te vrednosti. Interpretirajte rezultate in narišite sliko napovedanih verjetnosti glede na število let šolanja. Izračunajte razmerje obetov za posameznika z 20 leti šolanja glede na nekoga, ki ima 5 let šolanja manj!

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Educ	-,692	,134	26,809	1	,000	,500
	Age	,007	,003	4,993	1	,025	1,007
	Sex	,549	,096	32,477	1	,000	1,732
	educ2	,025	,005	27,656	1	,000	1,025
	Constant	3,458	,991	12,163	1	,000	31,743

a. Variable(s) entered on step 1: educ, age, sex, educ2.

PRIMER 2: TITANIK

Raziskovalno vprašanje: Napovedovanje preživetja potnikov na Titaniku na podlagi starosti, spola in potovalnega razreda.
Datoteka: Titanic.sav

1. Spremenljivke vstavite v model logistične regresije in podrobno interpretirajte rezultate. Za referenčno skupino izberite prvi razred.

Dependent: Survived=1, Died=0
Sex: Female=??, Male=??

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-,034	,006	29,510	1	,000	,966
	sex	-2,498	,166	226,304	1	,000	,082
	pclass			103,289	2	,000	
	pclass(Class 2 vs 1)	-1,281	,226	32,236	1	,000	,278
	pclass(Class 3 vs 1)	-2,290	,226	102,813	1	,000	,101
	Constant	3,522	,327	116,212	1	,000	33,855

a. Variable(s) entered on step 1: age, sex, pclass.

Pomen koeficientov
Sex -> log(OR) M vs F, v istem razredu, iste starosti
pclass (Class2 vs 1) -> log(OR) za potnika 2. razreda glede na potnika 1. razdreda, iste starosti, istega spola
pclass (Class3 vs 1) -> log(OR) za potnika 3. razreda glede na potnika 1. razdreda, iste starosti, istega spola

2. Narišite napovedane verjetnosti preživetja glede na starost.
(dodatno: v model poleg starosti vključite še kvadrirano vrednost starosti in komentirajte rezultate)
3. Izračunajte razmerje obetov za ženske glede na moške in podajte 95% interval zaupanja.
4. Izračunajte razmerje obetov za tretji razred glede na prvi razred in podajte 95% in 99% interval zaupanja.
5. Izračunajte razmerje obetov moškega v tretjem razredu glede na žensko v prvem razredu.
6. V model dodajte interakcijo med spolom in potovalnim razredom ter interpretirajte rezultate. Izračunajte razmerje obetov žensk glede na moške za vsak potovalni razred posebej!

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,038	,007	32,432	1	,000	,962
	Sex	-3,886	,493	62,235	1	,000	,021
	Pclass			91,786	2	,000	
	pclass(1)	-1,530	,567	7,288	1	,007	,217
	pclass(2)	-4,065	,511	63,301	1	,000	,017
	pclass * sex			43,913	2	,000	
	pclass(1) by sex	-,070	,631	,012	1	,911	,932
	pclass(2) by sex	2,489	,540	21,220	1	,000	12,047
	Constant	4,804	,547	77,091	1	,000	122,040

Pomen koeficientov
Sex -> log(OR) M vs F, v PRVEM razredu, iste starosti

pclass(1) -> log(OR) za zensko 2. razreda vs zensko 1. razreda, iste starosti
pclass(2) -> log(OR) za zensko 3. razreda vs zensko 1. razreda, iste starosti

Sex+pclass(1) by sex -> log(OR) M vs F, v 2. razredu, iste starosti
Sex+pclass(2) by sex -> log(OR) M vs F, v 3. razredu, iste starosti

pclass(1)+pclass(1) by sex -> log(OR) za moskega 2. razreda vs moskega 1. razreda, iste starosti
pclass(2)+pclass(2) by sex -> log(OR) za moskega 3. razreda vs moskega 1. razreda, iste starosti

Ocenjene verjetnosti (model brez interakcije vs model z interakcijo)
Z 22 let, 1. razred: 0.94 vs 0.98
M 22 let, 1. razdred, 0.57 vs 0.52

Z II, 22 let, 0.81 vs 0.92
M II, 22 let, 0.27 vs 0.17

Z III , 22 let, 0.61 vs 0.47
M III, 22 let, 0.11 vs 0.18

DODATNI PRIMERI

1. Pri prvem primeru: Kako bi se spreminjala RR in OR, ce vzorec volilcev ne bi bil slucajni, ampak bi namenoma vkljucili v raziskavo 10 krat vec volicev za Clintona kot za ostale kandidate?

	Female	Male	Tot
Clinton	10a	10b	10(a+b)
Bush or Perrot	c	d	c+d
Total	10a+c	10b+d	10(a+b) + c + d

Ali lahko iz tega vzorca ocenimo delez moskih volilcev (oziroma zenskih), ki so volili za Clintona?

Ali lahko iz tega vzorca ocenimo delez Clintonovih volilcev (oziroma za ostale kandidate), ki so zenskega spola?