Spørsmål

21 april 2004

Krav til semesteroppgåva

Spørsmål:

1. *er det et krav om at vi skal ha en dummykodet variabel med i oppgaven?*

   Svar: Det er eit krav at det skal vere med ein nominalskalavariabel med fleire enn 2 kategoriar. Denne kan inkluderast ved å dummykode den

Det er kanskje grunn for å repetere litt av kravspesifikasjonane?

Krav til semesteroppgåve (1)

- **FORORD**
- Dersom du nyttar **data frå SSB sine granskingar** skal følgjande med i ein fotnote eller i eit forord:
  - "(En del av) de data som er benyttet i denne publikasjonen er hentet fra .....undersøkelsen (årstall). Data i anonymisert form er stilt til disposisjon gjennom Norsk samfunnsvitenskapelig datatjeneste (NSD). Innsamling og tilrettelegging av data ble opprinnelig utført av Statistisk Sentralbyrå. Hverken Statistisk Sentralbyrå eller NSD er ansvarlige for analysen av dataene eller de tolkninger som er gjort her."
- For kommunedata skriv ein ikkje «i anonymisert form».

Krav til semesteroppgåve (2)

- **Innbinding** er ikkje nødvendig. Men om de ønskjer å binde inn oppgåva skal det nyttast innbinding i **A4 format**.
- **Tittelsida skal** minimum innehalde studentnummer og tittel som indikerer avhengig variabel. For somme formål kan namn vere nyttig, men det er frivillig om namnet skal stå på oppgåva.

Krav til semesteroppgåve (3)

- **KRAVSPESIFIKASJONAR**
- a) Med utgangspunkt i deskriptiv statistikk for variablane som skal inkluderast i modellen, skal fordelinga deira beskrivast og mogelege transformasjonar vurderast. **Transformasjonar skal takast i bruk dersom dette kan forbetre analysen substansielt** (dvs. det er teoretiske grunnar til å tru at det marginale sambandet mellom forklaringsvariabel og avhengig variabel er kurvelineært, jfr. pkt d) eller dersom det kan gjere testprosedyrane meir truverdige (residualen kjem nærmare opp mot normalfordelinga)

Krav til semesteroppgåve (4)

- b) Modellen skal innehalde minst ein kategorisk forklaringsvariabel med meir enn to kategoriar (MÅLENIVÅ: nominalskala).
- c) Det skal gjennomførast ei drøfting av mogelege interaksjonar og minst eitt interaksjonsledd skal testast.
- d) Mogelege kurvelineære samanhengar skal drøftast og minst ein kurvelineære samanheng skal testast.

Krav til semesteroppgåve (5)

Med utgangspunkt i den første modellen skal følgjande drøftast

- e) I OLS-regresjon skal normalfordelinga av feilleddet vurderast.
- f) I OLS-regresjon skal det testast for heteroskedastisitet.
- g) I OLS regresjon skal effekten av autokorrelasjon vurderast.
- h) I logit regresjon skal diskrimineringsproblem vurderast.

Krav til semesteroppgåve (6)

- i) I både OLS og logit-regresjon skal multikollinearitetsproblem vurderast
- j) I både OLS og logit-regresjon skal effekten av utliggarar og innflytelsesrike case vurderast og eventuelt illustrerast.
- k) I både OLS og logit-regresjon skal modellspesifikasjonen vurderast.

Dikotomisering og dummykoding (1)

Spørsmål: *hva er forskjellen på dikotomisering og dummykoding?*

Svar: Dikotomisering av ein variabel tyder å dele skalaen i 2 deler (eventuelt samle kategoriane i 2 grupper) slik at variabelen berre får 2 verdiar (høg/lav, mye\lite, pluss\minus, etc.). Alle variablar kan dikotomiserast. Når ein gjer det vil ein sjølvsagt tape mye informasjon. Håpet er at det ein taper ikkje er vesentleg i høve til det ein drøftar.

Dikotomisering og dummykoding (2)

- Dummykoding er ein spesiell måte å kode/ lage hjelpevariable på. For kategoriar i nominalskalavariable eller intervall i ordinal- eller intervallskalavariable lagar ein hjelpevariable som får verdien 1 dersom caset har ein verdi lik kategoriverdien eller ein verdi som ligg i intervallet på ordinal- eller intervallskalaen og 0 dersom det ikkje har ein verdi som spesifisert
- Ved hjelp av dummykoding av kvar kategori på ein nominalskalavariabel treng ein ikkje miste noko av informasjonen i variabelen

Dummykoding har mange namn

Forfattar	Dummy	Effect	Contrast	
Hamilton	Dummy	Effect		
Hardy	Dummy	Effect	Contrast	
Menard/ **SPSS**	Indicator/ Simple	Deviation		
Weisberg	Dummy/ Indicator			
Hosmer& Lemeshow	Reference Partial	Deviation Marginal		

SPSS har mange måtar å kode (1)

- **Indicator**
  - Contrasts indicate the presence or absence of category membership. The reference category is represented in the contrast matrix as a row of zeros.
- **Simple**
  - Each category of the predictor variable (except the reference category) is compared to the reference category.
- **Difference**
  - Each category of the predictor variable except the first category is compared to the average effect of previous categories. Also known as **reverse Helmert contrasts**.

SPSS har mange måtar å kode (2)

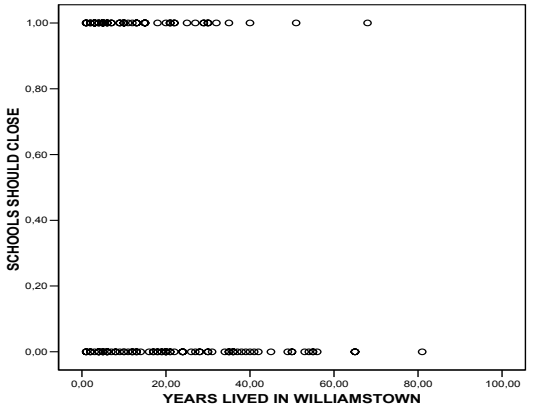
- **Helmert**
  - Each category of the predictor variable except the last category is compared to the average effect of subsequent categories.
- **Repeated**
  - Each category of the predictor variable except the first category is compared to the category that precedes it.
- **Deviation**
  - Each category of the predictor variable except the reference category is compared to the overall effect.
- **Polynomial**
  - Orthogonal polynomial contrasts. Categories are assumed to be equally spaced. Polynomial contrasts are available for numeric variables only.

Testing av linearitet i logiten

- **Spørsmål: Hvordan teste for linearitet i logiten? Det hadde vært fint med en ny gjennomgang, gjerne med en litt nærmere spesifisering av hvordan dette testes i SPSS.**
- **Svar**: Kurvelinearitet i logiten kan gi skeive parameterestimat. For å teste om Logiten er lineær i ein x-variabel kan vi gjere følgjande
  - Gruppere x-variabelen
  - For kvar gruppe finne y-gjennomsnitt og rekne det om til logit
  - Lag ein graf av logitane mot gruppert x

Statistiske problem: linearitet i logiten?

- Spreiingsplott for y-x er lite informative sidan y berre har to verdiar
- Y= Lukke skolen mot
- X= år budd i byen



Logiten

- L= naturleg logaritme (Oddsen for y=1) = **ln(p/(1-p))** der p=Pr{y=1}
- For å estimere p treng vi ei gruppe case der nokre har y=1 andre har y=0
- Dersom vi deler opp x-variabelen i intervall vil vi normalt for kvart intervall finne ei gruppe case som har y=1. For kvar gruppe slik definert kan vi rekne ut ein logit
- p = y-gjennomsnitt for dummykoda variable = prosenten/100 med verdien 1 på y

Eksempel

SCHOOLS SHOULD CLOSE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	OPEN	87	56,9	56,9	56,9
	CLOSE	66	43,1	43,1	100,0
	Total	**153**	**100,0**	**100,0**	

Descriptive Statistics

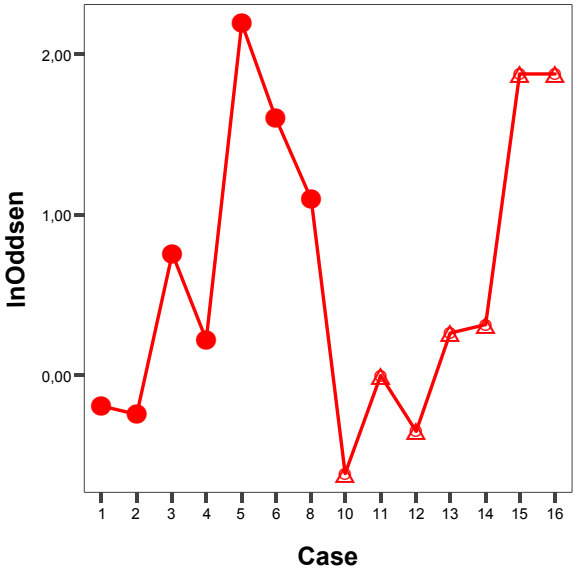
	YEARS LIVED IN WILLIAMSTOWN	SCHOOLS SHOULD CLOSE	Valid N (listwise)
N	153	153	153
Minimum	1,00	,00	
Maximum	81,00	1,00	
Mean	19,2680	,4314	
Std. Deviation	16,95466	,49689	

Count

		SCHOOLS SHOULD CLOSE	
		OPEN	CLOSE
YEARS LIVED IN WILLIAMSTOWN (Banded)	<= 3,00	7	13
	4,00 - 6,00	14	14
	7,00 - 11,00	7	10
	12,00 - 22,00	22	17
	23,00 - 33,00	11	8
	34,00 - 44,00	13	2
	45,00+	13	2
YEARS LIVED IN WILLIAMSTOWN (Banded)	1	24	29
	2	14	18
	3	17	8
	4	10	8
	5	9	1
	6	5	1
	7	4	
	8	3	1
	9	1	

		SCHOOLS SHOULD CLOSE				
		OPEN		CLOSE		Total
		Count	Row %	Count	Row %	Count
lived(Banded)	<= 3,00	7	35,0%	13	65,0%	**20**
	4,00 - 6,00	14	50,0%	14	50,0%	**28**
	7,00 - 11,00	7	41,2%	10	58,8%	**17**
	12,00 - 22,00	22	56,4%	17	43,6%	**39**
	23,00 - 33,00	11	57,9%	8	42,1%	**19**
	34,00 - 44,00	13	86,7%	2	13,3%	**15**
	45,00+	13	86,7%	2	13,3%	**15**
Group Total		87	56,9%	66	43,1%	153
Lived9	1-9	24	45,3%	29	54,7%	**53**
	10-18	14	43,8%	18	56,3%	**32**
	19-27	17	68,0%	8	32,0%	**25**
	28-36	10	55,6%	8	44,4%	**18**
	37-45	9	90,0%	1	10,0%	**10**
	46-54	5	83,3%	1	16,7%	**6**
	55-63	4	100,0%	0	0%	**4**
	64-72	3	75,0%	1	25,0%	**4**
	73-81	1	100,0%	0	0%	**1**
Group Total		87	56,9%	66	43,1%	153

Prosent med y=1 i gruppa av lived	ln(p/(100-p))
45,30	**-,19**
43,80	**-,25**
68,00	**,75**
55,60	**,22**
90,00	**2,20**
83,30	**1,61**
100,00	**missing**
75,00	**1,10**
100,00	**missing**
Gruppering brukt i forelesinga 17 mars:	
35,00	**-,62**
50,00	**,00**
41,20	**-,36**
56,40	**,26**
57,90	**,32**
86,70	**1,87**
86,70	**1,87**



Til venstre (7 første punkt) gir dagens oppdeling av lived. (2 punkt forsvinn sidan p=100)

Til høgre er forelesinga(17 mars) si oppdeling basert på

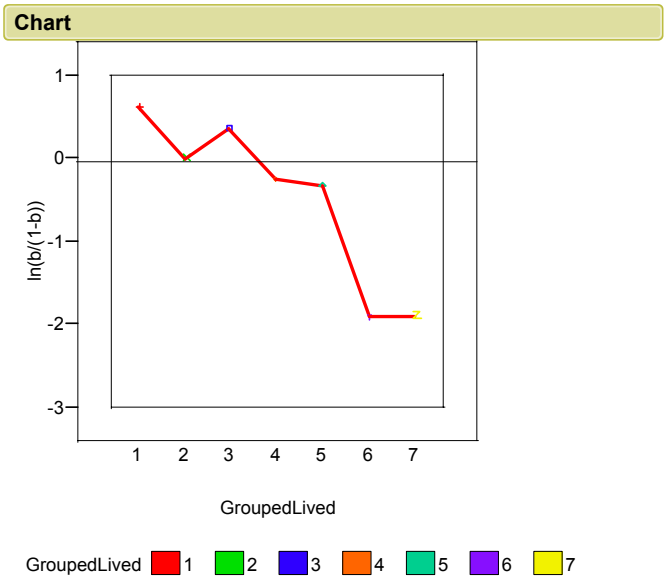
$p^F = 100-p$

Linearitet i logiten: eksempel

SCHOOLS SHOULD CLOSE		YEARS LIVED IN WILLIAMSTOWN (Banded)						
		<= 3	4-6	7-11	12-22	23-33	34-44	45+
N	OPEN	7	14	7	22	11	13	13
N	CLOSE	13	14	10	17	8	2	2
Within group	Mean (=p)	,65	,50	,59	,44	,42	,13	,13
Logit	Ln(p/(1-p))	0,619	0	0,364	-0,241	-0,323	-1,901	-1,901

Er logiten lineær i "år budd i byen"?

Tja, kanskje det.



Effekten av måleskala …

- Jeg holder på å teste ulike variabler i logistisk regresjon. De fleste er blitt signifikantet etter at jeg har lagt til samspillsvariabler og kvadratledd, problemet er at på en av variablene blir exp(B) veldig stor, over 15000 på det meste.
- Dette kan jo umulig stemme, men hvordan kan jeg gjøre noe med det. Den blir noe lavere når jeg fjerner kvadratleddet til variabelen, men kvadratleddet er sig så jeg må vel i grunnen ha det med??

Skalering med 1/1000

	B	S.E.	Wald	df	Sig.	Exp(B)
Educ	-,168	,080	4,414	1	,036	,845
Lived	-,052	,013	14,619	1	,000	,950
Constant	2,833	1,166	5,907	1	,015	17,004

	B	S.E.	Wald	df	Sig.	Exp(B)
Educ1000del	-167,984	79,955	4,414	1	,036	,000
Lived1000del	-51,615	13,500	14,619	1	,000	,000
Constant	2,833	1,166	5,907	1	,015	17,004

Bruk av kvadratledd

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	educ	,993	,614	2,616	1	,106	2,700
	EducSQ	-,044	,023	3,558	1	,059	,957
	lived	-,073	,036	4,200	1	,040	,930
	LivedSQ	,000	,001	,459	1	,498	1,000
	Constant	-4,428	3,993	1,230	1	,268	,012

a. Variable(s) entered on step 1: educ, EducSQ, lived, LivedSQ.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Educ1000del	993,076	613,943	2,616	1	,106	.
	Educ1000delSQ	-44048,2	23351,521	3,558	1	,059	,000
	Lived1000del	-73,092	35,665	4,200	1	,040	,000
	Lived1000delSQ	427,124	630,466	,459	1	,498	3,14+185
	Constant	-4,428	3,993	1,230	1	,268	,012

a. Variable(s) entered on step 1: Educ1000del, Educ1000delSQ, Lived1000del, Lived1000delSQ.

Formulering av modell

- **Spørsmål:** Formulering av en modell: Hvor mye skal egentlig være med, ref. slide 16 siste forelesning. Trodde det kun var populasjonsligningen, "La Y være... Sett X til... etc", og forutsetningen for OLS/ Logistisk regresjon. Det virker som det er mye mer omfattende i slide 16.
- **Svar:** Vel, ikkje så mye meir …

Formulering av modellar

- Definisjon av elementa i modellen
  - **variablar**, feilledd, populasjon og utval
- Definisjon av relasjonar mellom elementa
  - **likninga som bind elementa saman**, utvalsprosedyre, tidsrekkefølgje av hendingar og observasjonar,
- Presisering av føresetnader for bruk av gitt estimeringsmetode
  - tilhøve til substansteori (**spesifikasjon**)
  - **fordeling og eigenskapar ved feilledd**

Elementa i modellen

- **Variablar:** fenomenet vi ønskjer å studere må kunne observerast og seiast å ha ulike tilstandar eller uttrykksformer i ulike einingar i den populasjonen vi observerer. Vi må finne variasjon.
- **Feilledd:** feilleddet er ein abstrakt sekk som inneheld alle dei mange aspekta av populasjonen som vi ikkje er i stand til å observere og inkludere i modellen.
- **Populasjon:** kven eller kva er det vi ønskje å seie noko om?
- **Utval:** idealet er eit reint tilfeldig utval, om vi ikkje kan få det må vi vite nøyaktig korleis utvalsmetoden er knytt opp mot den avhengige variabelen (fenomenet) vi ønskjer å studere

Relasjonar mellom elementa

- **Likninga**: relasjonar mellom variablar
- **Utvalsprosedyre**: skeive (biased) utval pga seleksjon og manglande data
- Tidsrekkjefølgje av hendingar og observasjonar: kausal retning
- Samvariasjon, genuin/ spuriøs samvariasjon
  - Konklusjonar om kausalsamband krev genuin samvariasjon

Føresetnader for bruk av gitt estimeringsmetode

For å nytte OLS metoden til å estimere ein lineær modell må følgjande føresetnader gjerast:

- I. Modellen er korrekt, dvs.:
  - alle relevante variablar er med
  - ingen irrelevante er med
  - modellen er lineær i parametrane
- II. Gauss-Markov krava for «Best Linear Unbiased Estimates» (BLUE) er oppfylt
- III. Feilleddet er normalfordelt

Betinga effekt plott (1)

- **Spørsmål:**
  - Bør man alltid sette dummy-variablene til 0?

  Svar: nei, kva som skal setjast inn er avhengig av kva du er ute etter å finne svar på
  - Tester man da en gitt variabel på referanse-kategorien?

  Svar: (eg veit ikkje om eg skjønar spørsmålet)

  Men dersom alle inkluderte dummyvariablar er sett til null vil vi estimert verdi for referansegruppa

Betinga effekt plott (2)

– Kan man sette inn både mean og max/min i en slik ligning, eller må man holde seg til den ene eller den andre?

Svar: Dei variabelverdiane ein set inn for å lage eit betinga effekt plott er valde ut frå at dei skal illustrere situasjonen i den gruppa som vert definert av dei valde variabelverdiane. Om vi ikkje er interessert i spesielle grupper kan gjennomsnittspersonen vere eit startpunkt for granskinga