SOS3003 Eksamensoppgåver

Oppgåve 3 gitt hausten 2001

Erling Berge

Haust 2001 Oppgåve 3

I tabellvedlegget til oppgåve 3 er det estimert 7 ulike modellar av "Besøke husflidsforretning"

a) Lag eit konfidensintervall for effekten av "E.utdanning" i modell 1. Korleis kan ein tolke parameterestimatet for "E.utdanning"? Korleis tolkar vi den oppgitte oddsraten for "Kvinne"?

b) Formuler den modellen som er estimert i modell 2. Finn ut om "Ekteskapeleg status" gir eit signifikant bidrag til modellen. Bruk modell 3 til å finne forventa verdi av sannsynet for å vitje husflidsforretninga for ein ugift aleinebuande 50 år gammal mannleg universitetslærar frå Trondheim med 19 års utdanning.

c) Kva er definisjonen av Oddsen for å vitje husflidsforretning for den persontypen som er definert i pkt b)? Bruk definisjonen og modell 2 til å finne oddsraten for å velje å vitje husflidsforretning mellom ein mann med 19 års utdanning og ein med 18 års utdanning. Skriv opp formelen for å finne betinga effektplott for samanhengen mellom sannsyn og alder i modell 3.

d) Drøft om føresetnadene for modell 2 kan seiast å vere stetta. Drøft særleg problem med kurvelinearitet, multikollinearitet og diskriminering.

Haust 2001 Oppgåve 3a

- Lag eit konfidensintervall for effekten av ”E.utdanning” i modell 1.

I tabellvedlegget for oppgåve 3 modell 1 finn vi at

	Estimate	Std Error	Chi Square	Prob> ChiSq	Odds Ratio	VIF
Kvinne	1.2748	0.112	128.70	<.0001	3.578	1.008
E.utdanning	0.0255	0.0172	2.20	0.138	1.291	1.006

Til skilnad frå SPSS vil kolonnen med oddsratar her gi høvestalet mellom oddsen for å ha variabelen sin høgaste verdi og oddsen for å ha variabelen sin lågaste verdi.

Haust 2001 Oppgåve 3a

Eit 95% konfidensintervall for effekten av E.utdanning er da gitt ved

- $b_k - t_\alpha * SE(b_k) < \beta_k < b_k + t_\alpha * SE(b_k)$

- $0.0255 – 0.0172 * 1.96 < b_{E.utdanning} < 0.0255 + 0.0172 * 1.96$
- $0.0255 – 0.0337 < b_{E.utdanning} < 0.0255 + 0.0337$
- $-0.0082 < b_{E.utdanning} < 0.0592$

Haust 2001 Oppgåve 3a

- Korleis kan ein tolke parameterestimatet for "E.utdanning"?
- I 95 av 100 granskingar av spørsmålet om kven som ønskjer å vitje husflidsforretning vil konklusjonen at eitt år ekstra utdanning for personen gir ein tilvekst i logiten som er mellom –0.008 og 0.06 logiteiningar vere rett. Sidan 0 ligg i intervallet kan vi ikkje forkaste nullhypotesa om E.utdanning ikkje har nokon effekt på sannsynet for å vitje husflidsforretning.

Haust 2001 Oppgåve 3a

- Dersom vi reknar med den effekten som er estimert utan omsyn til at den "eigentleg" er 0, vil vi i kolonnen for oddsraten finne raten mellom oddsen i den høgaste utdanningskategorien i høve til oddsen i den lågaste utdanningskategorien, dvs. oddsen for å vitje husflidsforretning for personar med 17 års utdanning i høve til oddsen for personar med 7 års utdanning. Dei som er best utdanna har ein odds som er 1.2905474 gonger større enn dei som har lågast utdanning. Auken i oddsen for kvart år ekstra utdanning vert exp[bE.utdanning] = exp[0.02550665] = 1.02583473, eller omlag 2,6 % auke for kvart ekstra år med utdanning. Med 10 år meir utdanning (17år – 7år) vert auken i oddsen lik exp[0.02550665*10] = 1.2905474.

Haust 2001 Oppgåve 3a

- Koeffesienten for utdanning kan og tolkast i samband med sannsynet for at Y=1. Da må vi i tillegg ta omsyn til kva verdi dei to andre variablane i likninga har. Vi finn sannsynet ut frå samanhengen P=1/(1+exp(-L)) der P er sannsynet for eit case med logit L. Meir spesifikt finn vi i modell 1 at for case i er
- $Pr[Y\ i = 1 \mid X_{1i}, X_{2i}, X_{3i}] =$

  $1/ (1+ Exp[- \{-2.713 + 1.275*Kvinne_i + 0.0255*E.utdanning_i + 0.0667*Barn\_i\_hushaldet_i\}])$
- Samanhengen mellom utdanning og sannsyn for vitje husflidsforretning studerer vi best ved hjelp av betinga effekt plott.

Haust 2004 © Erling Berge 7

Haust 2001 Oppgåve 3a

- Korleis tolkar vi den oppgitte oddsraten for "Kvinne"?
- Oddsraten for Kvinne er lik 3.578. Det tyder at oddsen for å vitje husflidsforretning er meir enn 3 og ein halv gong større for kvinner enn for menn.

Haust 2004 © Erling Berge 8

Haust 2001 Oppgåve 3b

- Formuler den modellen som er estimert i modell 2.
- Finn ut om ”Ekteskapeleg status” gir eit signifikant bidrag til modellen.
- Bruk modell 3 til å finne forventa verdi av sannsynet for å vitje husflidsforretninga for ein ugift aleinebuande 50 år gammal mannleg universitetslærar frå Trondheim med 19 års utdanning.

Haust 2001 Oppgåve 3b
Formuler den modellen som er estimert i modell 2.

- Når vi skal formulere ein modell må vi
  - definere elementa som inngår i modellen (variablar og data),
  - definere relasjonane mellom elementa (regresjonslikninga), og
  - presisere kva føresetnader som ein må gjere for å bruke modellen.

Haust 2001 Oppgåve 3b
Formuler den modellen som er estimert i modell 2.

Vari-abel	Variabelnamn	Kommentar
Y	Besøke husflidsforretning	Y=1 dersom person i ønskjer å vitje lokalt kunstgalleri, elles er Y=0
X_1	Kvinne	dummykoda
X_2	E.utdanning	år
X_3	Barn i hushaldet	1 = ja , 0 = nei
X_4	Alder	år

I eit tilfeldig utval på 2948 personar frå den norske befolkninga frå 1991 er det opplysningar om desse variablane. Vi lar indeksen i=1,2, ... ,2948 indikere kva for ein person opplysningane gjeld for.

Haust 2001 Oppgåve 3b
Formuler den modellen som er estimert i modell 2

- I populasjonen føreset vi at det er eit logistisk samband mellom sannsynet for å ha verdien Y=1 på den avhengige variabelen og dei uavhengige X-variablane. Modell 2 er da definert ved at vi lar

  $Pr[Y_i=1] = E[Y_i]$, der $Y_i=1/(1+\exp\{-L_i^*\}) + e_i$,

  der e_i er feilleddet, L_i^* er estimert forventa verdi av logiten, L_i der i = 1,2,3, … ,2948, og logiten er definert ved
- $E[L_i]=b_0 +b_1X_{1i} +b_2X_{2i} +b_3X_{3i} +b_4X_{4i}$

Haust 2001 Oppgåve 3b
Formuler den modellen som er estimert i modell 2

- Ein føreset vidare at modellen er rett spesifisert, dvs.:
  - den funksjonelle forma for alle betinga sannsyn for Y=1 er logistiske funksjonar av X-ane (dette svarar til at Logiten er lineær i parametrane)
  - ingen relevante variablar er utelatne
  - ingen irrelevante variablar er inkluderte
- alle X-variablane er utan målefeil
- alle case er uavhengige
- det er ikkje perfekt multikollinearitet

og kanskje også

- det er ikkje perfekt diskriminering og
- utvalet er sort nok

Dei siste punkta er ikkje teke med som føresetnadar av Hamilton (1992, jfr. side 225 og 233) men representerer substansielt sett same type problem som multikollinearitet. Ein bør vidare vere merksam på at innverknadsrike case, høg grad av multikollinearitet og sterk grad av diskriminering fører til problem for estimeringa i form av upresise estimat (stor varians).

Haust 2004 © Erling Berge 13

Haust 2001 Oppgåve 3b
Er ”Ekteskapeleg status” signifikant?

- to modellar kan samanliknast ved å nytte den kjikvadratfordelte testobservatoren
- $\chi^2_H = -2[\mathcal{LL}(\text{K-H=liten mod.}) - \mathcal{LL}(\text{K=stor mod.})]$
- der $\mathcal{LL}$ står for logLikelihooden, K er talet på parametrar i den største modellen og H = talet på fridomsgrader for testen (= talet på variablar som skil mellom dei to modellane = skilnaden i talet på estimerte parametrar). I dette høvet er H = 3, talet av inkluderte dummyvariablar for “Ekteskapeleg status”)

Haust 2004 © Erling Berge 14

Haust 2001 Oppgåve 3b
Er ”Ekteskapeleg status” signifikant?

Liten Modell	Log-Likelihood	Stor Modell	Log-Likelihood
1	-1257.924	4	-1227.913
2	-1220.659	5	-1211.427
3	-1199.139	6	-1197.048

- χ^2_H = -2[$\mathcal{LL}$(K-H=liten mod.) - $\mathcal{LL}$(K=stor mod.)]
- Finn kjikvadratverdien for testen av modell 6 mot 3

Haust 2001 Oppgåve 3b
Verdi av sannsynet

- Bruk modell 3 til å finne forventa verdi av sannsynet for å vitje husflidsforretninga for ein ugift aleinebuande 50 år gammal mannleg universitetslærar frå Trondheim med 19 års utdanning.
- I variabelen E.utdanning vil ein person med 19 års utdanning få verdien 17. Vi kan da anten nytte 17 eller 19 i utrekninga av forventa verdi av logiten:

X_1	Kvinne =0
X_2	E.Utdanning = 19
X_3	Barn i hushaldet = 0
X_4	Alder = 50

Haust 2001 Oppgåve 3b
Verdi av sannsynet

Variabelnamn	Verdi av variabel	Parameterestimate	Variabelverdi * Parameterestimat
Konstant		-6.5102512	-6.5102512
Kvinne	0	1.34401309	0
E.utdanning	19	0.05900844	1.12116036
Barn i hushaldet	0	0.25424529	0
Alder	50	0.13641411	6.8207055
Alder*Alder	50*50	-0.0011803	-2.95075
		Logitverdi	**-1.5191353**

Sannsynet finn vi da som Pr(Y=1 | x-verdiar i oppgåveteksten) =
1/(1+exp[-Logitverdi]) = 1/(1+exp[-(**-1.5191353**)]) = **0.17958889**

Haust 2004 © Erling Berge 17

Haust 2001 Oppgåve 3c

- Kva er definisjonen av Oddsen for å vitje husflidsforretning for den persontypen som er definert i pkt b)?
- Bruk definisjonen og modell 2 til å finne oddsraten for å velje å vitje husflidsforretning mellom ein mann med 19 års utdanning og ein med 18 års utdanning.
- Skriv opp formelen for å finne betinga effektplott for samanhengen mellom sannsyn og alder i modell 3

Haust 2004 © Erling Berge 18

Haust 2001 Oppgåve 3c
Definisjonen av Oddsen

- Oddsen er definert som sannsynet for å vitje husflidsforretning dividert med ein minus sannsynet for å vitje husflidsforretning. Logiten er definert som den naturlege logaritmen til oddsen. Dermed vil vi finne oddsen ved å opphøgje grunntalet e i Logiten; dvs. O_i = exp{L(i)}, der i= person av typen definert i pkt b.

Haust 2004 © Erling Berge 19

Haust 2001 Oppgåve 3c
oddsraten mellom menn med 19 og 18 års utdanning

- Oddsraten finn vi som høvetalet mellom to Odds. La j= person av typen "i" men med variabelverdien x-1 i staden for x. Da er Oddsraten (i-person i høve til j-person på x-variabelen) = Oi / Oj = exp[L(i)] / exp[L(j)] = exp[L(i)-L(j)]
- Dersom to personar, i og j, har same variabelverdiar med unntak av at den eine har 19 års utdanning (i) og den andre 18 (j), vil differansen mellom logitane deira i dette høvet bli:
- L(i)-L(j) = 0.0754151*E.utdanning(i) - 0.0754151*E.utdanning(j) = 0.0754151* (E.utdanning(i) - E.utdanning(j)) = 0.0754151* (19-18) = 0.0754151.

Haust 2004 © Erling Berge 20

Haust 2001 Oppgåve 3c
oddsraten mellom menn med 19 og 18 års utdanning

- Dermed blir oddsraten mellom dei to personane
- O_i / O_j = exp[0.0754151] = 1.07833167
- Med andre ord: oddsen for å vitje husflidsforretning aukar med omlag 8% for kvart år ekstra udanning om alt anna er likt.
- Dvs: oddsraten O_i / O_j = exp[$b_{E.utdanning}$]

Haust 2004 © Erling Berge 21

Haust 2001 Oppgåve 3c
Betinga effekt plott

Forventa verdi av logiten er i modell 3 estimert til

- L(i) = -6.510 + 1.344*Kvinne + 0.0590*E.utdanning + 0.254*Barn i hushaldet + 0.136*Alder - 0.00118*Alder*Alder
- Sannsynet finn vi som
- Pr(Y=1) = 1/(1 + Exp{-L(i)}) = 1/(1 + Exp{-(-6.510 + 1.344*Kvinne + 0.0590*E.utdanning + 0.254*Barn i hushaldet + 0.136*Alder - 0.00118*Alder*Alder)})
- For å få eit betinga effektplott av samanhengen mellom alder og sannsyn må vi sette in verdiar av variablane Kvinne, E.utdanning og Barn i hushaldet.

Haust 2004 © Erling Berge 22

Haust 2001 Oppgåve 3d

- Drøft om føresetnadene for modell 2 kan seiast å vere stetta. Drøft særleg problem med kurvelinearitet, multikollinearitet og diskriminering.

Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

Krava til modellen er definert under punkt b.

- Vi kan ikkje sjekke om variablane er utan målefeil eller om utvalet er av uavhengige case. Vi vil ta dette for gitt for dette utvalet.
- Spesifikasjonskravet kan vi derimot seie ein del om.
  - Modellen 2 har ikkje irrelevante variablar. Alle inkluderte variablar er signifikante med ein p-verdi mindre enn 0.0001.
  - Utelatne relevante variablar kan vi ikkje seie noko om

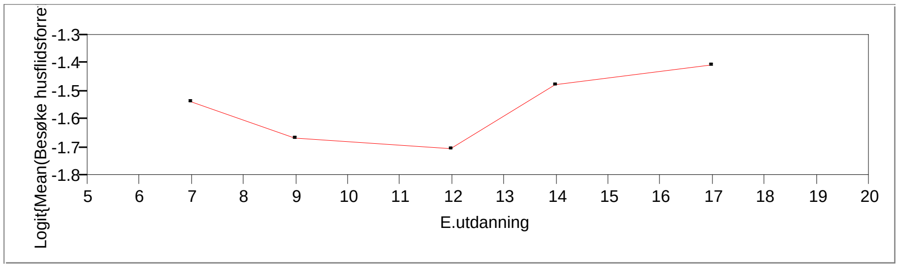
Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

- Kravet om at logiten skal vere lineær i parametrane kan vi sjekke. Av dei fire variablane er to dummykoda og kan ikkje vere kurvelineære. Dei to andre variablane, Alder og E.utdanning, kan vi sjekke om dei eigentleg er kurvelineære ved hjelp av tabellane av logiten til gjennomsnittleg verdi av "Besøke husflidsforretning" etter alder og utdanningsnivå.
- Alder er tydelegvis sterkt kurvelineær medan det for E.utdanning berre er visse veike tendensar.

Haust 2004 © Erling Berge 25

Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

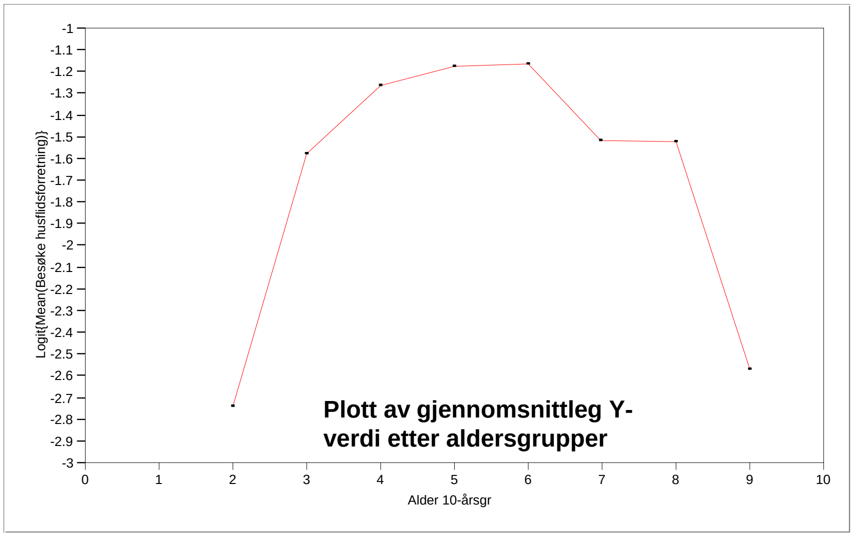
Plott av gjennomsnittleg Y-verdi etter utdanningsgrupper



- Slike figurar er svært sensitive for korleis skalaen på y-aksen er framstilt. Vi må derfor sjå på variasjonsbreidda for den observerte logiten.

Haust 2004 © Erling Berge 26

Haust 2001 Oppgåve 3d
Føresetnadene for modell 2



Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

- Spørsmålet om kurvelinearitet i logiten kan også svarast på ved å teste om andregradspolynom med Alder og E.utdanning gir signifikante bidrag til modellformuleringa. For dei 7 modellane har vi følgjande loglikelihoodar:

Nr	Liten Modell u/ Ektesk st.	LogLikelihood	Nr	Stor Modell m/Ektesk st.	LogLikelihood
1	– u/ alder	-1257.924	4	– u/ alder	-1227.913
2	– m/alder	-1220.659	5	– m/alder	-1211.427
3	– m/alder*alder	-1199.139	6	– m/alder*alder	-1197.048
7	– m/E.utd*E.utd	1199.115			

Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

- Modell 6 mot 4:
- χ^2_H = -2* ([-1227.913494]-[-1197.048429]) = -2*(-30.865065)= 61.73013
- Testen har 2 fridomsgrader (H=2) og nullhypotesa om ingen effekt av alder vert klart forkasta. Testen er likevel overflødig så lenge begge dei to ledda i alderspolynomet er så tydeleg signifikante kvar for seg.

Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

- Modellestimering krev vidare at det ikkje er **perfekt** multikollinearitet eller **perfekt** diskriminering. I og med at modellane 1-7 faktisk har latt seg estimere viser dette at krava er oppfylt
- Vi skal likevel undersøkje om der er stor grad av multikollinearitet og diskriminering sidan dei begge kan gje store standardfeil og upresise parameterestimat.

Haust 2001 Oppgåve 3d
Føresetnadene for modell 2

Multikollinearitet: stor VIF?

- VIF er stor berre der vi introduserer andregradsledd

Diskriminering: nullceller i krysstabellar?

- Null berre for dummykoda variabel ”Uoppgitt e.status”. Der denne dummyen er bruk vert resultatet merka ”unstable”