EKSAMENSOPPGÅVER SVSOS316
VÅR 2001
FRAMLEGG TIL LØYSING

Erling Berge
Institutt for sosiologi og statsvitenskap
Norges Teknisk Naturvitskapelege Universitet

«Bruksanvisning»

Når ein går igang med å løyse oppgåver må ein ha i minnet at oppgåvene ofte er problematiske i høve til modellbygginga sitt krav om at modellen må vere fundert på den best tilgjengelege teorien. Mangelen på teoretisk fundament for oppgåvene kan forsvarast ut frå to perspektiv. Det avgjerande er rett og slett mangelen på tid og høvelege data for å lage eksamensoppgåver av den «realistiske» typen det er tale om her. Men tar ein for gitt at oppgåvene sjeldan kan seiast å vere teoretisk velfundert, gir jo dette studentane lettare gode poeng i arbeidet med å vurdere modellane kritisk ut frå spesifikasjonskravet.

Når ein studerer framlegga til løysingar er det viktig å vere klar over at det som er presentert ikkje er nokon fasit. Dei fleste oppgåvene kan løysast på mange måtar. Dei tekniske sidene av oppgåvene er sjølvsagt eintydige. Men i dei mange vurderingane (som t.d. «Er denne residualen tilstrekkeleg nær normalfordelinga til at vi kan tru på testane?») er det nett vurderingane og argumentasjonen som er det sentrale.

På eksamen er tida knapp. Svært få rekk i eksamenssituasjonen å gjere grundig arbeid på heile oppgåvesettet. I arbeidet med dette løysingsframlegget har det vore gjort meir arbeid enn det som ein ventar å finne til eksamen. Somme stader er det teke med meir detaljar i utrekningar og tilleggsstoff som kan vere relevant, men ikkje nødvendig. Men det er ikkje gjort like grundig alle stader.

Det må takast atterhald om feil og lite gjennomtenkte vurderingar. Underteikna har like stor kapasitet til å gjere feil som andre. Kritisk lesning av studentar er den beste kvalitetskontroll ein kan ønskje seg. Den som finn feil eller som meiner andre vurderingar vil vere betre, er hermed oppfordra til å seie frå
(t.d. på e-mail: <Erling.Berge@sv.ntnu.no>)

OPPGÅVE 1 (vekt 0,1)
a) Forklar kva glatting (smoothing) er.

Når ein måler samme fenomenet med samme metoden gong etter gong får vi ein tidsserie (t.d. med n målingar). Endringane i måleresultat kan forklarast dels som tilfeldige påverknader, mellom anna målefeil, og dels som reultat av substansielle endringar å årsaksfaktorar.

Dei tilfeldige feila kan ofte minskast ved glatting, særleg dersom måltidpunkta kjem ofte i høve til endringstakten i dei substansielle årsaksfaktorane. Glatting kan gjerast ved å ta eit gjennomsnitt av nærliggande målingar, t.d. tre målingar som kjem etter kvarandre. Ved å flytte seg ei måling oppover i serien for kvar gong ein tar gjennomsnittet vil vi få like mange gjennomsnitt som vi har målingar minus 2:

$e_t^* = 1/3(e_{t-1} + e_t + e_{t+1})$ $\quad t=2, \ldots, n-1$

Glatting kan også gjerast ved å nytte andre transformasjonar av data, t.d. ved å vekte kvar observasjon i glattingsfunksjonen ulikt og ved å ta omsyn til fleire enn 3 punkt i gongen (jfr. Hamilton 1992:121-123)

b) Korleis kan ein nytte dummyvariablar til å teste for kurvesamanhengar?

Dersom vi har mistanke om at ein intervallskalavariabel ikkje har lineær verknad på den avhengige variabelen kan vi sjekke om dette er tilfelle ved hjelp av dummyvariablar. Vi deler da opp den uavhengige variabelen i passeleg lange segment (intervall) og kodar om kvart intervall til ein dummyvariabel. Vi erstattar så intervallskalavariabelen med gruppa av dummyvariablar i regresjonen og estimerer effekten av kvart intervall i høve til referanseintervallet. Plottar vi desse effektane i eit diagram med avhengig og uavhengig variabel (intervallskalaform) vil vi sjå om effektane er jamnt stigande (minkande) eller om det er former for kurvesamanhengar.
(jfr. Hardy 1993:78-80)

Vi kan handsame ein ordinalskalavariabel på samme måten. Mangelen på avstand mellom kategoriane gjer det sjølvsagt vanskeleg å seie noko om kva matematisk form ein eventuell kurvesamanheng får. Men for å nytte ein ordinalskala direkte i ein regresjon må det vere linearitet i dei avstandane som implisitt vert antatt å ligge i skalaen når den vert brukt slik. Med dummykoding kan dette sjekkast.

OPPGÅVE 2

(OLS-regresjon, vekt 0,45)

I tabellvedlegget til oppgåve 2 er det estimert 8 modellar av eiga inntekt, somme med estimerte verdiar for manglande inntektsopplysningar andre med utelating av personar der opplysningar mangla.

a) Bruk modell 1 for å finne eit konfidensintervall for effekten av å ha heiltidsarbeid. Vurder om det er ein lineær eller kurvelineær samanheng mellom alder og inntekt. Finn ut frå modell 3 forventa inntekt for ei 40 år gammal kvinne med 12 års utdanning og heiltidsarbeid ved NTNU.

I modell 1 er $b_{Heiltidsarbeid}$, effekten av Heiltidsarbeid, oppgitt til å vere 98.818511 med ein standardfeil på 2.499229. Dersom vi kan gå ut frå at feilledda er normalfordelte vil eit 95% konfidensintervall (5% signifikansnivå) vere gitt ved

$$b_{Heiltidsarbeid} - SE_{Heiltidsarbeid} * t_{2,5\%} < \beta_{\mathbf{Heiltidsarbeid}} < b_{Heiltidsarbeid} + SE_{Heiltidsarbeid} * t_{2,5\%}$$

der b er den estimerte regresjonskoeffesienten, SE er standardfeilen til regresjonskoeffesienten og t er fraktilen i t-fordelinga i ein tosidig test med signifikansnivå 0,05. I følge tabell A4.1 hos Hamilton (1992:350) vil vi med meir enn 120 fridomsgrader ha at $t_{2,5\%}$ =1,96 (einsidig test, 5% i tosidig test). Set vi inn i formelen finn vi no at

$$98.818511 - 2.499229 * 1.96 < \beta_{Heiltidsarbeid} < 98.818511 + 2.499229 * 1.96$$

dvs.

$$93.92002216 < \beta_{\mathbf{Heiltidsarbeid}} < 103.71699984$$

I modell 1 er alder ein lineær faktor i forklaringa av E.inntekt. I modell 2 er alder og alder*alder begge like klart signifikante faktorar for forklaringa av E.inntekt. I modell 2 er dessutan justert determinasjonskoeffesient auka med 4 prosentpoeng frå 0.513 til 0.553. Det kan utan tvil konkluderast med at alder viser ein kurvelineær samanheng med forventa eiga inntekt.

I oppgåveteksten finn vi ein person som har følgande variabelverdiar
Alder = 40
Mann = 0
E.utdanning = 12
Heiltidsarbeid = 1
Offentleg sektor = 1

Vi skal finne forventa inntekt for denne personen ut frå følgande estimerte samanheng mellom variablane og E.inntekt :

E.innt(EM) =			=			=		
	-68.01107			-68.01107			-68.01107	
	+4.3638167	Alder		+4.3638167	*40		+4.3638167	*40
	-105.888	Mann		-105.888	*0		0	
	+5.5702413	E.utdanning		+5.5702413	*12		+5.5702413	*12
	+66.897826	Heiltidsarbeid		+66.897826	*1		+66.897826	
	+9.1808397	Offentleg sektor		+9.1808397	*1		+9.1808397	
	-0.041206	Alder*Alder		-0.041206	*40*40		-0.041206	*40*40
	+4.5283472	Alder*Mann		+4.5283472	*40*0		0	
	-0.040529	Alder*Alder*Mann		-0.040529	*40*40*0		0	
	+3.0419769	E.utdanning*Mann		+3.0419769	*12*0		0	
	+29.348031	Heiltidsarbeid*Mann		+29.348031	*1*0		0	
	-30.07997	Offentleg sektor*Mann		-30.07997	*1*0		0	

Forventa verdi av E.innt(EM) =
-68.01107 +4.3638167*40 +5.5702413*12 +66.897826 +9.1808397 -0.041206*1600 =
-68.01107 +174.552668 +66.8428956 +66.897826 +9.1808397 –65.9296 =
183.5335593
eller omtrent 183.500 kroner

b) Formuler den modellen som er estimert som Modell 4. Vurder om testane i modell 4 er truverdige. Test om bustad gir ei signifikant yting til å forklare variasjonen i inntekt.

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

I modell 4 er følgande variablar definert:

Y = E.inntekt m/ estimert missing (E.innt(EM))
X_1 = Alder
X_2 = Mann
X_3 = E.utdanning
X_4 = Heiltidsarbeid
X_5 = Offentleg sektor
X_6 = Alder*Alder
X_7 = Mann* Alder
X_8 = Mann* Alder*Alder
X_9 = Mann*E.utdanning
X_{10} = Mann*Heiltidsarbeid
X_{11} = Mann*Offentleg sektor
X_{12} = Bost. Sentrum storby
X_{13} = Bost. Forst. storby
X_{14} = Bost. Småby
X_{15} = Bost. Tettst.
X_{16} = Bost. Uoppg
X_{17} = Funksjonær
X_{18} = Sjølvstendig
X_{19} = Elev/ student
X_{20} = Pensjon/ trygd
X_{21} = Andre KtL
X_{22} = Uoppgitt KtL

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

I populasjonen antar vi at det er eit lineært eller kurvelineært samband mellom den avhengige variabelen, Y, og dei uavhengige X-variablane. Dette tyder i vårt høve at

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} + \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \beta_{12} X_{12i} + \beta_{13} X_{13i} + \beta_{14} X_{14i} + \beta_{15} X_{15i} + \beta_{16} X_{16i} + \beta_{17} X_{17i} + \beta_{18} X_{18i} + \beta_{19} X_{19i} + \beta_{20} X_{20i} + \beta_{21} X_{21i} + \beta_{22} X_{22i} + \varepsilon_i$$

når vi lar i gå over heile populasjonen. Lar vi k=0, 1, 2, ... ,22, vil β_k vere dei ukjente parametrane som viser kor mange måleeiningar av Y vi får i tillegg ved å auke X_k med ei måleeining. ε_i er eit feilledd som fangar opp dei faktorane vi ikkje har observert saman med reint tilfeldig støy i målinga av Y_i.

Vi kan estimere dei ukjente parametrane i denne modellen dersom vi har observasjonar for eit reint tilfeldig utval frå populasjonen og vi kan gjere følgjande føresetnader:

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, dvs.:
- Faste x-verdiar (dvs. vi kan i prinsippet trekke nye utval med samme x-verdiar men ulik y-verdi).
- Feilledda har forventning 0 for alle i, dvs: $E(\varepsilon_i) = 0$ for alle i .
- Feilledda har konstant varians (homoskedastisitet) dvs: $var(\varepsilon_i) = \sigma^2$ for alle i .
- Feilledda er ukorrelerte med kvarandre (ikkje autokorrelasjon)
  dvs: $cov(\varepsilon_i, \varepsilon_j) = 0$ for alle $i \neq j$.

III. Normalfordeling av feilleddet:
- Feilledda er normalfordelte med samme varians for alle case,
  dvs: $\varepsilon_i \sim N(0, \sigma^2)$ for alle i .

Generelt veit vi at tekniske og substansielle problem som t.d. ikkjelineær samband, utelatte variablar, målefeil i dei uavhengige variablane, heteroskedastisitet, autokorrelasjon og ikkje-normalfordelte feilledd vil alle føre til at t- of F-testane ikkje blir truverdige. Autokorrelasjon er ikkje aktuelt i eit enkelt tilfeldig utval av personar. Av dei andre faktorane som kan påverke truverdet til testane er det berre heteroskedastisitet og ikkje-normalfordelte residualar som kan undersøkjast.

Vi kan svare på spørsmålet i oppgåva på to måtar. Dersom vi vil avvise at testane er truverdige treng vi berre påvise svikt i ein av føresetnadene. Dersom vi ikkje meiner at ein eventuell svikt i nokon av føresetnadene er alvorleg, kan vi konkludere med at testane er truverdige.

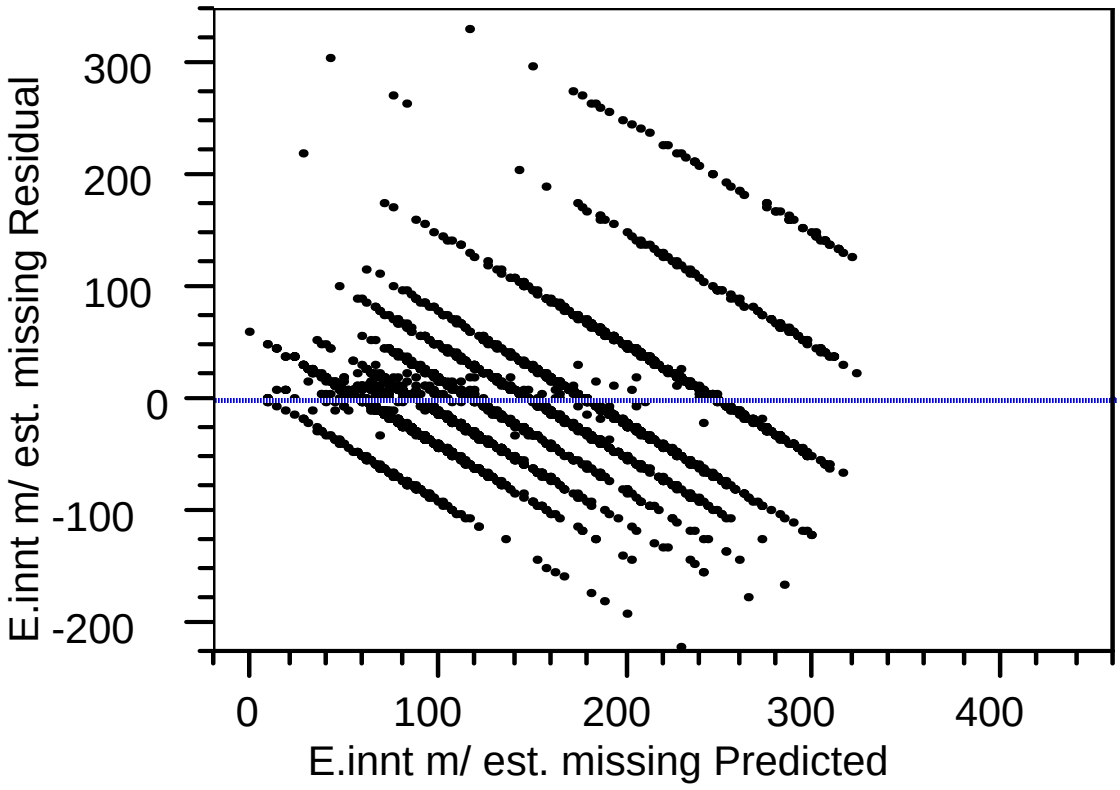
Plottet av residualen mot predikert verdi av E.inntekt i modell 4 har ei tydeleg (men ikkje svært sterk) ”diamantform”. Dette indikerer eit visst element av heteroskedastistet ut over det som er bygd inn gjennom kategoriseringa av inntektsvariabelen (jfr

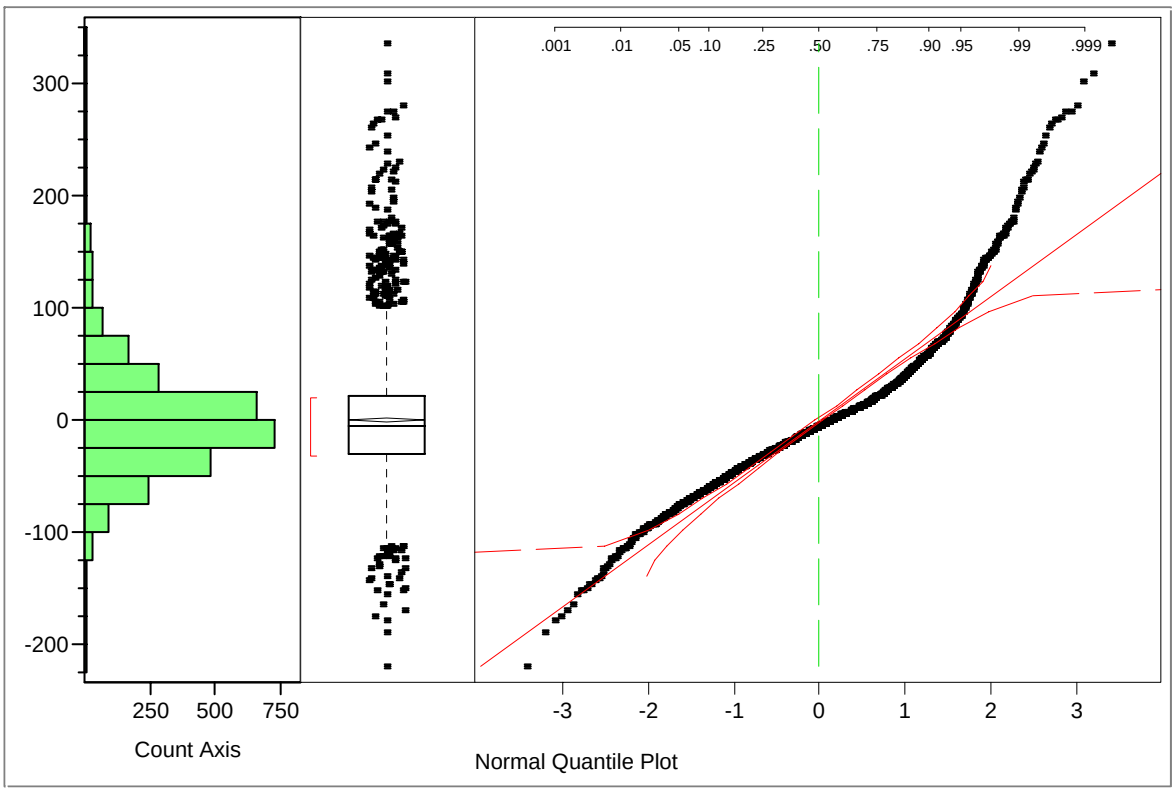
variabeldefinisjonen som har 8 kategoriar, pluss ein ekstra lagt inn i samband med estimeringa av missing).

Dette bør gjere at vi ser på testane med ein rimeleg grad av skepsis. Truverdet er ikkje stort.

Fordelinga av residualen viser at den er ein del høgreskeiv med fleire høge verdiar. Medianen til residualane er på –4.63, litt mindre enn gjennomsnittet. Men med variasjonsbreidda –218 til 336.37 er ikkje dette mye.
Kvartilavviket, IQR = 21.89 - (-31.06) = 52.95. IQR/1.35 skal i ein normalfordeling vere omlag lik standardfeilen (standardavviket, St.dev.). Vi ser her at standardfeilen = 55.40 medan IQR/1.35 = 39.22. Heller ikkje dette kan kallast eit stort avvik gitt variasjonsbreidda til residualen.

Ser vi på kvantil-normal plottet ser vi da også at for hoveddelen av variasjonsområdet er fordelinga rimeleg nær normalfordelinga. Basert berre på fordelinga av residualane kan vi konkludere med at vi nok kan ha noko tillit til testane. Sett saman med heteroskedastisitetproblemet bør vi vere noko meir forsiktig med å lite på konklusjonane.





Vi skal teste om **"Bustad"** yter signifikant til å forklare variasjonen i inntekt. Vi må da gå ut frå at føresetnadene som er presisert ovanfor er rette.

Bustad er dummykoda. Kategorien «**Spredtbygd**» er utelaten og fungerer som referansekategori. Dei fem andre kategoriane i variabelen er inkludert i modell 4, men ikkje i modell 5. Vi kan da nytte testobservatoren F på modellane 4 (stor modell) og 5 (liten modell) for å samanlikne to «nesta» modellar.

Analysis of variance

Modell 4

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	22	14126182	642099	207.6285
Residual	2925	9045674	3093	**Prob > F**
C. Total	2947	23171856		0.0000

Modell 5

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	17	14017080	824534	263.8934
Residual	2930	9154776	3124	**Prob > F**
C. Total	2947	23171856		0.0000

Når vi samanliknar to modellar estimert på samme utval av n case, ein modell med K parametrer og ein med K - H parametrar vil observatoren

$$F^{H}_{n-K} = \frac{(RSS[K-H] - RSS[K]) / H}{RSS[K] / (n-K)}$$

vere F-fordelt med H og (n-K) fridomsgrader dersom det faktisk er rett at dei H ekstra variablane ikkje har effekt (dersom H_0 er rett). RSS(*) er residualane sin kvadratsum i dei ulike modellane. Vi forkastar null-hypotesa om at alle koeffesientane til dei H ekstra variablane er null med signifikansnivået α dersom F^{H}_{n-K} er større en α-fraktilen i F-fordelinga med H og (n-K) fridomsgrader.

Samanliknar vi modell 4 og 5 ser vi at
H = 5
K = 23
n-K= 2948 - 23 = 2925
RSS[K-H] = 9154776
RSS[K] = 9045674
Vi finn da at F^{H}_{n-K} = 7.05582248487,
og sidan 5 % fraktilen(α) i F-fordelinga med 5 og 2925 fridomsgrader er 2.21 (Hamilton 1992, tabell A4.2) vil vi forkaste nullhypotesa om at variabelen «Bustad» ikkje yter til å forklare variasjonen i eiga inntekt. Den forklarte variansen aukar likevel berre frå 60.5% i modell 5 til 61% i modell 4.

c) Modell 6 er identisk med modell 4 men er estimert utan den personen som har størst innverknad på estimatet av modell 4. Kva kan seiast om denne personen? Kva konsekvensar har det for regresjonsresultatet at personen vert utelaten?

Opplysningar om Case 1537

Variabel	**Kode**	**I oppgåva**	**Kommentar**
Næring	3	Manglar def	3=samferdsel/transport/post/tele
Ekteskapeleg status	1	Manglar def	1=gift
Heiltidsarbeid	0		0=deltid, varierer, nei, missing
Offentleg sektor	1		Uvanleg inntekt, næring, sektor?
Talet på arbeidstakarar i husstanden	2	Manglar def	
Barn i hushaldet (ja/nei)	1	Manglar def	
Mor's utdanning	9	Manglar def	
Far's utdanning	7	Manglar def	
E.utdanning	12		
Alder	42		
Mann	1		
E.inntekt	450		
HH.(hushalds)inntekt	180	Manglar def	**Må vere feil!**
Bustadstype	5		5=spredt
Kjelde til Livsopphald	12		

Ut frå dei opplyningane vi har her ser vi at den personen som har stor innverknad er ein 42 år gammal mann. Han bur i spredtbygd strok, er gift, har barn og arbeider deltid (ikkje heiltidsarbeid) i offentleg sektor innan næringane samferdsel/ transport/ post/ tele. Kjelde til livsopphald er imidlertid uoppgitt. Også kona har arbeid. Foreldra har lav utdanning og personen har sjølv vidaregåande skole. Denne personen oppgir så å ha ei inntekt i øverste intervallet (450.000). Medan Hushaldsinntekta er oppgitt å vere 180.000. Dette siste må i tilfelle vere feil. Det kan vel vere tale om ei mistyding der det er konas inntekt som er oppgitt.

Dei mest uvanlege variabelverdiane er vel likevel at vedkommande arbeider i offentleg sektor og er utan heiltidsarbeid saman med den høge inntekta og at vedkommande bur spredbygd, der relativt få offentlege tilsette tener så mye. Høg inntekt i offentleg sektor vil normalt kreve heiltidsarbeid og ein noko meir "urban" bustad.. Utanom offentleg sektor kan ein imidlertid tene bra på noko variabel arbeidsinnsats (t.d. fiske). Det er imidlertid ikkje umogeleg å tenkje seg karrierer som kan ha slike kombinasjonar av variabelverdiar. Dei er berre uvanlege.

Konsekvensar av å utelate person 1537

Samanlikning av modell 4 (med case 1537) og 6 (utan case 1537)

Term	Estimate model 4	Estimate model 6
Intercept	-25.75671	-25.07622
Alder	2.7856639	2.7568453
Mann	**-98.63631**	**-97.03559**
E.utdanning	4.4888481	4.4516546
Heiltidsarbeid	47.809716	47.546914
Offentleg sektor	7.0106699	6.9827594
Alder*Alder	-0.026519	-0.026278
Alder*Mann	4.6389816	4.5035116
Alder*Alder*Mann	-0.042205	-0.040677
E.utdanning*Mann	2.5075154	2.5174164
Heiltidsarbeid*Mann	**29.007274**	**30.471599**
Offentleg sektor*Mann	**-26.08115**	**-27.21207**
Sentrum storby	15.885219	16.676846
Forstad storby	20.133127	20.907522
Småby	8.8718665	9.6634437
Tettstad	8.8304113	9.5787324
Uoppg bostad	4.8355522	5.4611336
Funksjonær	33.852976	33.986939
Sjølvstendig	24.907939	24.969538
Elev/ student	-23.62938	-23.84837
Pensjon/ trygd	-5.323715	-5.282188
Andre KtL	-15.57144	-15.4818
Uoppgitt KtL	**-17.46346**	**-24.72841**

Substansielt ser vi at å fjerne denne personen gir ein auke i effekten på 6-800 kr for bustadsdummyane, 12-1500 for interaksjonsledda Heiltidsarbeid*Mann og Offentleg sektor*Mann. Sørste auken finn vi for Uoppgitt KtL. Men denne kategorien er i seg sjølv teoretisk uinteressant. Det er imidlertid verdt å merke seg at den personen vi fjernar har Uoppgitt KtL. For Mann finn vi ein nedgang i effekt på ca 1600.

Relativt finn vi at auken er på 41.6% for Uoppgitt KtL, 12.9% for Uoppgitt bostad, 8.9% for Småby, 8.5% for Tettstad, 5% for H.arb*Mann og for Sentrum storby, den er 4.3% for O.sek*Mann, og for Mann har vi ein nedgang på 1.6%. For dei andre er endringane mindre.

Den substansielt viktigaste verknaden finn vi dermed for Bustadsvariabelen der fjerning av eitt case frå referansekategorien fører til ein auke i inntektsskilnadene mellom dei som bur spreidt og dei andre kategoriane på 5-9%.

d) Manglande opplysnigar i variabelen "E.inntekt m/ est. missing" er erstatta med eit estimat frå Modell 7. Modell 7 er identisk med modell 3 men er estimert på faktiske observasjonar. Drøft generelt problemet med manglande observasjonar på avhengig variabel. Vurder konkret og substanisielt skilnadene mellom Modell 3 og 7 i estimerte effektar for "Mann" og "Offentleg sektor".

Manglande observasjonar

Når personar som skulle vore med i ei gransking ikkje er til stades eller ikkje vil svare på eitt eller fleire spørsmål står vi i fare for å få eit skeivt utval. Det foregår ein seleksjonsprosess. Generelt kan ein seie at dersom seleksjonen på nokon måte kan knyttast til den avhengige variabelen, vil estimata vi finn i utvalet ikkje kunne gjerast gjeldande for populasjonen. Estimata blir skeive. Spesielt vil dette gjelde for manglande svar på den avhengige variabelen.

Allment kan ein tenkje seg fleire ulike situasjonar:

1. Det manglar opplysningar for ei gruppe personar på ein eller fleire uavhengige variablar, eller det manglar opplysningar for dei personane som har visse gitte verdiar på ein uavhengig variabel, t.d. X>x' eller X<x'' (X-variabelen seiest da å vere trunkert).
   Dette er rekna som lite problematisk. Fråfallet gir mindre utvalsstorleik med dei problem det kan gi i form av mindre presisjon (større varians) i estimeringa, men fører elles ikkje til skeive parameterestimat.

2. Det manglar opplysningar om kva verdiar ei gruppe personar har på den avhengige variabelen. Dette fører ikkje til problem dersom personane er tilfeldig fordelt over variasjonsområdet til variabelen. Er dei ikkje det, vil det nærme seg situasjonen for sensurerte, selekterte eller trunkerte utval.

3. **Sensurerte utval.** Det manglar opplysningar om verdien på Y-variabelen for personar som har visse gitte verdiar på den avhengige variabelen, t.d. Y>y' eller Y<y'' (Y-variabelen seiest vere trunkert). Opplysningar om kva verdiar dei observerte personane har på X-variablane manglar ikkje. Utvalet vert i dette høvet kalla sensurert. Dette gir alvorlege problem. Parametrar vil bli skeivt estimert og modellen kan bli feilspesifisert.

4. **Selekterte utval.** Det manglar opplysningar om verdien på variabelen Y for personar som har visse gitte verdiar på ein uobservert variabel Z, t.d. Z<z' eller Z>z'' (Y-variabelen er også i dette høvet trunkert). Utvalet vert kalla selektert. Dersom Z på nokon måte er korrelert med den avhengige variabelen Y fører dette til problem. Parametrar vil bli skeivt estimert og modellen kan bli feilspesifisert.

5. **Trunkerte utval.** Det manglar opplysningar om verdiane på både y, og x-variablane for personar som har visse gitte verdiar på den avhengige variabelen, t.d. Y>y' eller Y<y'' (Y-variabelen er trunkert). Sidan også opplysningane om verdiar på x-variablane manglar vert heile utvalet kalla trunkert.
   Dette gir alvorlege problem. Parametrar vil bli skeivt estimert og modellen kan bli feilspesifisert.

Vi ser i modell 7 at det manglar inntektsopplysningar for 314 personar. Utvalet er ikkje selektert, men ligg ut frå vedlagte tabellar i grenselandet mellom å vere eit sensurert utval og eit utval der dei utelatte er tilfeldig fordelt over inntektsskalaen. Problemet oppstår ved at personar som faktisk er utan inntekter eller som har svært små og meir tilfeldige inntekter lar vere å svare på spørsmålet i mye større grad enn dei som har regulære inntekter.

Kva konsekvensar har det å erstatte manglande opplysningar med estimat basert på dei som har opplysningar?

Skilnader mellom modell 3 og 7 i estimerte effektar

Samanliknar vi dei betinga effektplotta for modell 3 og 7 synest dei fortelje samme historia. Det er ikkje stor skilnad mellom dei to modellane i kva effektar ein får estimert.

Term		**Estimate model 3**	**Estimate model 7**
Intercept		-68.01107	-67.00947
Alder	A	4.3638167	4.8758099
Mann	**M**	**-105.888**	**-98.59718**
E.utdanning	E.utd	5.5702413	5.3362693
Heiltidsarbeid	**H.arb**	**66.897826**	**62.605666**
Offentleg sektor	**O.sek**	**9.1808397**	**3.5888226**
Alder*Alder	A*A	-0.041206	-0.04714
Alder*Mann	A*M	4.5283472	3.9804485
Alder*Alder*Mann	A*A*M	-0.040529	-0.034699
E.utdanning*Mann	E.utd*M	3.0419769	3.0099967
Heiltidsarbeid*Mann	H.arb*M	29.348031	30.257213
Offentleg sektor*Mann	**O.sek*M**	**-30.07997**	**-22.35092**

Gjennomgåande ser det ut til at modell 7 har lågare absoluttverdi på effektane enn modell 3. Somme av skilnadene i koeffesientar ser ut til å vere store. Det gjeld særleg for effekten av Offentleg sektor og interaksjonen Offentleg sektor og Mann, til dels ser dei ut til å vere store også for Heiltidsarbeid og Mann.Vi finn at skilnaden mellom modellane 3 og 7 i effektane for Alder og Mann, og interaksjonane mellom mann og heiltidsarbeid og offentleg sektor er negativ. For dei andre er skilnadene i effekt positiv. Vurderingane av einskildkoeffesientar er imidlertid vanskeleg sidan pluss og minuseffektar i sum tenderer til å kansellere kvarandre. Vi skal vurdere meir konkret korleis det slår ut for effekten av Mann og Offentleg sektor:

Modell 3
E[E.innt(EM)] = -68.01107 + 4.3638167*A -105.888*M + 5.5702413*E.utd + 66.897826*H.arb + 9.1808397*O.sek -0.041206*A*A + 4.5283472*A*M - 0.040529*A*A*M + 3.0419769*E.utd*M + 29.348031*H.arb*M -30.07997*O.sek*M

Modell 7
E[E.innt] = -67.00947 + 4.8758099*A -98.59718*M + 5.3362693*E.utd + 62.605666*H.arb + 3.5888226*O.sek -0.04714*A*A + 3.9804485*A*M - 0.034699*A*A*M + 3.0099967*E.utd*M + 30.257213*H.arb*M -22.35092*O.sek*M

Ser vi no spesielt på **effekten av å vere Mann** finn vi
Modell 3:
-105.888*M + 4.5283472*A*M -0.040529*A*A*M + 3.0419769*E.utd*M + 29.348031*H.arb*M -30.07997*O.sek*M
Modell 7:
-98.59718*M + 3.9804485*A*M -0.034699*A*A*M + 3.0099967*E.utd*M + 30.257213*H.arb*M -22.35092*O.sek*M

Forskjellen mellom å vere Mann i modell 3 og Mann i modell 7 er dermed
(-105.888+98.59718)*M + (4.5283472-3.9804485)*A*M +
(-0.040529+0.034699)*A*A*M + (3.0419769-3.0099967)*E.utd*M +
(29.348031-30.257213)*H.arb*M + (-30.07997+22.35092)*O.sek*M

Set vi inn M=1 er dette tilnærma lik
-7.3 + 0.5*A - 0.006*A*A +0.03*E.utd - 1.0*H.arb - 7.7*O.sek

For unge menn på sei 20 år utanfor offentleg sektor, utan heiltidsarbeid og med t.d. 12 års utdanning finn vi da
-7.3 + 0.5*20 - 0.006*400 +0.03*12 = 0.66, dvs vi finn 660 kroner høgare inntekt i modell 3 enn i modell 7. For 30-åringar aukar skilnaden til 2660, for 40-åringar til 3460 og for 50 åringar er den på 3060
Å erstatte missing med estimerte verdiar på den måten vi har gjort det her, har i praksis lite å seie for effekten av å vere mann.

Ser vi på **effekten av å vere i Offentleg sektor** finn vi
Modell 3
9.1808397*O.sek -30.07997*O.sek*M
Modell 7
3.5888226*O.sek -22.35092*O.sek*M

Forskjellen mellom å vere i offentleg sektor i modell 3 og i modell 7 blir dermed
(9.1808397-3.5888226)*O.sek + (-30.07997+22.35092)*O.sek*M

Set vi inn O.sek=1 finn vi dette tilnærma likt 5.6 – 7.7*M

I offentleg sektor vil kvinner i følge modell 3 ha ei inntekt på ca 5600 kroner over det modell 7 gir, medan vi for menn finn ei inntekt som er ca 2100 kroner lågare.
Her må skilnaden seiast å ha substansiell betydning.

Ser vi på tabellane nedanfor ser vi også at mellom dei som vi ikkje har registrert inntekt for, er dei fleste ikkje i offentleg sektor og ikkje registrert med heiltidsarbeid. Dei er også gjennomgåande yngre og med noko lågare utdanning (meir konsentrert om 9-12 år).

Frekvensfordelingar for 314 personar utan inntektsopplysningar

Heiltidsarbeid

Level	Count	Prob
0	275	0.87580
1	39	0.12420

Offentleg sektor

Level	Count	Prob
0	283	0.90127
1	31	0.09873

E.utdanning

År	Count	Prob
7	69	0.21975
9	95	0.30255
12	126	0.40127
14	16	0.05096
17	8	0.02548

Alder 10-Årsgr

År	Count	Prob
-29	140	0.44586
30-39	56	0.17834
40-49	29	0.09236
50-59	29	0.09236
60-69	18	0.05732
70-79	29	0.09236
80-89	11	0.03503
90+	2	0.00637

Kjelde til livsopphald

Level	Count	Prob
1	31	0.09873
2	25	0.07962
3	7	0.02229
4	3	0.00955
5	4	0.01274
6=elev/stud/læ	94	0.29936
7	22	0.07006
8	27	0.08599
9	24	0.07643
10=gift uten a	41	0.13057
11	10	0.03185
12	26	0.08280

HH.inntekt (1000)

1000 kr	Count	Prob
60	2	0.00637
90	16	0.05096
120	5	0.01592
150	18	0.05732
180	30	0.09554
250	27	0.08599
350	18	0.05732
450	13	0.04140
998=miss	164	0.52229
999=miss	21	0.06688

Når vi erstattar missing med eit regresjonsestimat plasserer vi dei på den linja som er definert av dei som ikkje har missing. Så lenge missing fordeler seg tilfeldig ut over inntektsvariabelen vil ikkje erstatning av missing ha noko å seie.

Det synest ut frå tabellane ovanfor rimeleg å konkludere med at ein stor del av dei utelatte høyrer heime i grupper utan inntekt eller med låg inntekt. Dei er ikkje tilfeldig fordelt på inntektsskalaen. Det synest rimeleg å konkludere med at ein uobservert variabel, z, ar avgjerande for om vi har opplysning om inntekt. Utvalet vår er selektert.

Seleksjonsvariabelen er tilsynelatande sterkt korrelert med variablar som Offentleg sektor, Heiltidsarbeid og Alder. Der korrelasjonen mellom inntekta og determinantar for seleksjonskriteriet (som t.d. Heiltidsarbeid og Offentleg sektor) er sterk får vi store skilnader i parameterestimat. I ein tostegsmodell for for selekterte utval (Breen 1996:33-47) vil det vere rimeleg å freiste modellere z ved hjelp av desse variablane.

OPPGÅVE 3 (Logistisk regresjon, vekt 0,45)

I vedlagte tabellar er det estimert 4 ulike modellar av "Besøke lokalt kunstgalleri"

a) Lag eit konfidensintervall for effekten av "Mors utdanning" i modell 1. Korleis tolkar ein parameterestimatet for "Mors utdanning".

I tabellvedlegget for oppgåve 3 modell 1 finn vi at

	b	St. feil	ChiSq	Pr > ChiSq	oddsraten	VIF
Mors utdanning	0.11441208	0.0225642	25.71	<.0001	2.4975161	1.2684774

I modell 1 er effekten av Mors utdanning estimert til å vere 0.11441208 med ein standardfeil på 0.0225642. I logistisk regresjon er storleiken $t= b_k/ Se_{bk}$ tilnærma normalfordelt i store utval og i store utval er normalfordelinga og t-fordelinga tilnærma ekvivalente. I store utval kan vi med andre ord finne konfidensintervall for ein parameter i ein logistisk regresjonsmodell på samme måten som i OLS regresjon. Da vil eit 95% konfidensintervall vere gitt ved

$$b_{Kvinne} - SE_{Kvinne}*t_{2,5\%} < \beta_{Kvinne} < b_{Kvinne} + SE_{Kvinne}*t_{2,5\%}$$

der b er regresjonskoeffesienten, SE er standardfeilen til regresjonskoeffesienten og t er fraktilen i t-fordelinga i ein tosidig test med signifikansnivå 0,05. I følge tabell A4.1 hos Hamilton (1992:350) vil vi med meir enn 120 fridomsgrader ha at $t_{2,5\%}$ =1,96. Set vi inn i formelen finn vi no at

$$0.11441208 - 0.0225642* 1.96 < \beta_{Mors\ utdanning} < 0.11441208 + 0.0225642* 1.96$$
$$0.11441208 – 0.044225832 < \beta_{Mors\ utdanning} < 0.11441208 + 0.044225832$$
$$0.070186248 < \beta_{Mors\ utdanning} < 0.158637912$$

I 95 av 100 granskingar av spørsmålet om kven som ønskjer å vitje det lokale kunstgalleriet vil vi vente å finne at eitt år ekstra utdanning for personen si mor gir ein tilvekst i logiten som er mellom 0.07 og 0.16 logiteiningar.

Parameterestimatet for Mors utdanning kan tolkast på fleire måtar. Som nemnt ovanfor kan det sjåast som eit lineært tillegg i logiten for kvart år ekstra utdanning mor har. Det kan også fortelje oss kva oddsraten for å velje å vitje lokalt kunstgalleri er mellom ulike utdanningsgrupper. I kolonna for oddsraten finn vi raten for den høgaste utdanningskategorien i høve til den lågaste kategorien, dvs. For mødre med 15 års utdanning i høve til mødre med 7 års utdanning. Dei som har best utdannia mødre har ein odds som er 2.4975161 gonger større enn dei som har lågast utdanna mødre. Auken i oddsen for kvart år ekstra utdanning mor har vert exp[0.11441208] = 1.12121405959, eller omlag 12% årleg tilvekst. Over 8 år (=15-7) vert det exp[0.11441208*8] = 2.4975161

b) Formuler den modellen som er estimert i modell 3. Finn ut om "Bustadstype" gir eit signifikant bidrag til modellen. Vurder om føresetnadene for modellen kan seiast å vere stetta.

Formulere modell

La Y_i=1 dersom person i svarar at han eller ho ønskjer å vitje lokalt kunstgalleri og la Y_i=0 for alle andre svar.

La vidare

X_{1i}=	Kvinne,	dummy for kvinne,
X_{2i}=	E.utdanning	eiga utdanning i år,
X_{3i}=	Mors utdanning	Mors utdanning i år
X_{4i}=	Alder,	alder i år,
X_{5i}=	E.utdanning*Alder	interaksjonsledd
X_{6i}=	Alder**2,	Alder*Alder,
X_{7i}=	E.utdanning*Alder*Alder	interaksjonsledd

der i=1,2, … ,N gir identiteten til personane i populasjonen.

Modell 3 er da definert ved at vi antar at observasjonane våre kan modellerast ved å sette $Pr[Y_i=1] = E[Y_i]$, der $Y_i=1/(1+\exp\{-L_i^*\}) + \varepsilon_i$, ε_i er feilleddet, og L_i^* er estimert forvente verdi av logiten, L_i , som er modellert ved

$$E[L_i]=\beta_0 +\beta_1X_{1i} +\beta_2X_{2i} +\beta_3X_{3i} +\beta_4X_{4i} +\beta_5X_{5i} +\beta_6X_{6i} +\beta_7X_{7i}$$

Ein antar vidare at

- modellen er rett spesifisert, dvs.:
  - den funksjonelle forma for alle betinga sannsyn for Y=1 er logistiske funksjonar av X-ane (eller Logiten er lineær i parametrane)
  - ingen relevante variablar er utelatt
  - ingen irrelevante variablar er inkluder
- alle X-variablane er utan målefeil
- alle case er uavhengige
- fravær av perfekt multikollinearitet

og kanskje også

- fravær av perfekt diskriminering

Dette punktet er ikkje tatt med som forutsetning av Hamilton (1992 jfr side 225 og 233) men representerer substansielt samme type problem som multikollinearitet.

Ein bør vidare vere merksam på at innflytelsesrike case, høg grad av multikollinearitet og sterk grad av diskriminering fører til problem for estimeringa.

Teste "Bustadstype"

Dei fire modellane av «Besøke lokalt kunstgalleri» er hierarkisk oppbygd. "Bustadstyper" er inkludert berre i modell 4. Vi kan teste om "Bustadstype" bidrar signifikant til modellspesifikasjonen ved å samanlikne modell 4 med den nest "største" modellen, 3, i ein sannsynsratetest (Likelihoodrate test). Testen nyttar den kjikvadratfordelte testobservatoren

$$\chi^2_{\mathbf{H}} = -2\{\log_e L_{\mathbf{K-H}} - \log_e L_{\mathbf{K}}\}$$

der L står for Likelihooden, K er talet på parametrar i den største modellen (her modell 4) og H= talet på fridomsgrader for testen (= talet på variablar som skil mellom dei to modellane = skilnaden i talet på estimerte parametrar: her er dette 5, ein for kvar av dei inkluderte dummyvariablane for "Bustadstype").

Testen er basert på nullhypotesa at regresjonskoeffesientane for dei inkluderte dummyvariablane for bustad ikkje er statistisk ulik 0. Dersom denne hypotesa er rett er det urimeleg å vente at χ^2_H skal få ein verdi som er svært ulik 0.

Loglikelihooden ($\log_e L$) i modell 3 er -1076.740051 og i modell 4 er den -1066.775823 slik at

$$\chi^2_{\mathbf{H}} = -2([-1076.740051]-[-\ 1066.775823]) = -2*(-9.964228) = 19.928456$$

Med 5 fridomsgrader vil eit kjikvadrat på 11,07 eller større gi eit signifikansnivå (= sjansen for å forkaste ei rett nullhypotese) på 0.05 eller lågare. Med eit kjikvadrat på 19.928 vil vi derfor forkaste nullhypotesa.

Vurdering av føresetnader

I kravet om at modellen er rett spesifisert, dvs.:

- den funksjonelle forma for alle betinga sannsyn for Y=1 er logistiske funksjonar av X-ane (eller Logiten er lineær i parametrane)
- ingen relevante variablar er utelatt
- ingen irrelevante variablar er inkluder

kan vi sjekke om alle inkluderte variablar er relevante og vi kan i prinsippet sjekke funksjonsforma sjølv om vi her manglar opplysningar for å kunne gjere det. Det vi ikkje kan sjekke er om alle relevante variablar er med. Men ser vi på tabellen nedanfor er einaste rimelege konklusjon at modellen ikkje er rett spesifisert.

Count Total % Col % Row %	Most likely Y=1	0	
Observed Y=1	4 0.14 26.67 1.02	389 13.20 13.26 98.98	393 13.33
0	11 0.37 73.33 0.43	2544 86.30 86.74 99.57	2555 86.67
	15 0.51	2933 99.49	2948

Tabellen viser oss observerte verdiar av Y=1 samanlikna med estimat der Pr(Y=1)>=0.5 gir predikert (most likely) Y = 1. Vi ser at av 393 case med observert Y=1 vil modellen predikere rett for berre 4 stykker. Mindre enn ca 1% er predikert rett for dei som har Y=1.. At det så vert predikert rett verdi for 99.57 % av dei som har observert Y=0 er berre ein refleks av at det vert predikert Y=0 for 99.5% av alle case.

Ein optimist vil legge merke til at 86.44% av alle case er predikert rett. Den kritiske ser med ein gong at ein ville gjort det betre om alle hadde vorte gitt prediksjonen Y=0, det ville gitt 86.67% rett prediksjon.

Konlusjonen må bli at modellen reflekterer svært dårleg faktisk tidsbruk. Det må opplagt mangle relevante variablar.

I estimatet av modellen finn vi også at fleire av variablane i modellen ikkje bidrar signifikant. Men ser vi på rekkja av modellar 1-4 ser vi at dette rimelegvis har å gjere med multikollinearitet introdusert ved andregradsledd og samspelsledd. I modell 1 er alle einskildvariablane signifikante og med normale VIF verdiar. I modell tre har vi toleranseverdiar (=1/VIF) på ned mot 0.0022 som er langt under normale faregrenser.

Kravet om at alle X-variablane er utan målefeil og at alle case er uavhengige kan vi ikkje seie noko om.

Og krava om fravær av perfekt multikollinearitet og fravær av perfekt diskriminering er oppfylt så lenge modellen teknisk lar seg estimere. Men som notert ovanfor er graden av multikollinearitet alt i modell 3 så høg at det er på grensa til det som teknisk lar seg gjere. Diskriminering har vi ikkje data til å undesøkje.

Konklusjonen må bli at det kan reisast rimeleg tvil om krava til ein logistisk regresjonsmodell er stetta i modell 3.

c) Bruk modell 4 til å finne forventa verdi av sannsynet for å vitje det lokale kunstgalleriet for ein 50 år gammal mannleg universitetslærar frå Trondheim med 19 års utdanning når du også får vite at mor hans hadde 8 års utdanning. Skriv opp formelen for å finne betinga effektplott for samanhengen mellom sannsyn og alder.

Forventa verdi av sannsyn

Ut frå teksten kan vi sette opp følgande variabelverdiar
Kvinne = 0
E.utdanning = 19
alternativt kan det nyttast verdien 17 ut frå kodereglane for variabelen
Mors utdanning = 8
alternativt kan det nyttast verdien 7 ut frå kodereglane for variabelen
Alder = 50
Småby=1 (alle andre bustadsvariablar = 0)
alternativt kan ein akseptere Sentrum storby = 1 (og alle andre bustadsvariablar = 0)
Vurderinga at ein skal erstatte faktiske opplysningar med kodar nytta i regresjonsestimatet er grei, men innan rimelege grenser bør det også vere lov å ekstrapolere innan eit rimeleg variasjonsområde for det koda variabelintervallet.

Sidan bustad er inkludert mellom opplysningane er det rimeleg å nytte modell 4 til å finne verdien av Logiten for å vitje lokalt kunstgalleri

-2.0989542		-2.0989542		-2.0989542	-2.0989542
0.26397597	Kvinne	0.26397597	*0	0	
-0.247483	E.utdanning	-0.247483	*19	-0.247483*19	-4.702177
0.11756567	Mors utdanning	0.11756567	*8	0.11756567*8	0.94052536
-0.1388798	Alder	-0.1388798	*50	-0.1388798*50	-6.94399
0.0173218	E.utdanning*Alder	0.0173218	*19*50	0.0173218*19*50	16.45571
0.00140234	Alder*Alder	0.00140234	*50*50	0.00140234*50*50	3.50585
-0.0001657	Alder*Alder*E.utdanning	-0.0001657	*50*50*19	-0.0001657*50*50*19	-7.87075
0.75375453	Sentrum storby	0.75375453	*0	0	
0.56692945	Forstad storby	0.56692945	*0	0	
0.52327448	Småby	0.52327448	*1	0.52327448	0.52327448
0.13174341	Tettstad	0.13174341	*0	0	
0.18032952	Uoppg bostad	0.18032952	*0	0	
				L(i)	**-0.19051136**

Oddsen for å vitje lokalt kunstgalleri for personen i oppgåveteksten slik det er rekna i tabellen vert exp[L(person i oppgåvetekst)] = exp[-0.19051136] = 0.826536368223
Forventa verdi av sannsynet for å vitje det lokale kunstgalleriet for ein 50 år gammal mannleg universitetslærar frå Trondheim med 19 års utdanning når vi også veit at mor hans hadde 8 års utdanning vert tilsvarande.
Pr(Y=1 | x-verdiar i oppgåveteksten, inkl Trondheim=småby) = 1/(1+exp[-L(i)]) = 1/(1+exp[0.19051136]) = **0.4525**

Ved å velje andre verdiar for dei tre diskutable variablane finn vi andre verdiar av det forventa sannsynet:

1. Dersom vi set Trondheim = sentrum storby,
   Mors utdanning=8 og E.utdaninng=19 får vi L(i) = 0.03996869
   Pr(Y=1) = 0.5100
2. Dersom vi nyttar Trondheim = småby,
   Mors utdanning=7 og E.utdanning=17 får vi L(i) = -0.71679103
   Pr(Y=1) = 0.3281
3. Dersom vi nyttar Trondheim = sentrum storby,
   Mors utdanning=7 og E.utdanning=17 får vi L(i) = -0.48631098
   Pr(Y=1) = 0.3808

Betinga effekt plott

NB: Her må vi gi verdiar til variablane Kvinne, E.utdanning, Mors utdanning og Småby. Alle "lovlege verdiar er her gangbart. Eksempelvis kan vi setje Kvinne=0, E.utdanning=17, Mors utdanning=7, Småby=1.

Eit betinga effekt plott illustrerer kva ei likning med gitte koeffesientar tyder. Plottet kan ikkje takast som prov på noko som helst, verken kurvelinearitet eller interaksjonar. Men der vi har etablert signifikante samanhengar vil plottet gi oss eit bilete av kva samanhengane faktisk tyder.

Formelen for å finne eit betinga effektplott for samanhengen alder og forventa sannsyn er i dette høvet

P(alder) =

E[Pr{Y=1| Alder, Kvinne=0, E.utdanning=17, Mors utdanning=7, Småby=1}] =

1/(1+exp[-L(Alder | Kvinne=0, E.utdanning=17, Mors utdanning=7, Småby=1)])

der

L(Alder | Kvinne=0, E.utdanning=17, Mors utdanning=7, Småby=1) =

-2.0989542 +0.26397597*Kvinne -0.247483*E.utdanning +0.11756567*Mors utdanning -0.1388798*Alder +0.0173218*E.utdanning*Alder +0.00140234*Alder*Alder -0.0001657*Alder*Alder*E.utdanning +0.75375453*Sentrum storby +0.56692945*Forstad storby +0.52327448*Småby +0.13174341*Tettstad +0.18032952*Uoppg bostad =

-2.0989542 +0.26397597*0 -0.247483*17 +0.11756567*7 -0.1388798*Alder +0.0173218*17*Alder +0.00140234*Alder*Alder -0.0001657*Alder*Alder*17 +0.75375453*0 +0.56692945*0 +0.52327448*1 +0.13174341*0+0.18032952*0 =

-2.0989542 -0.247483*17 +0.11756567*7 -0.1388798*Alder +0.0173218*17*Alder +0.00140234*Alder*Alder -0.0001657*Alder*Alder*17 +0.52327448*1 =

-4.95993103+ (-0.1388798 +0.0173218*17)*Alder

+(0.00140234 -0.0001657*17)*Alder*Alder =

-4.95993103+ 0.1555908*Alder –0.00141456*Alder*Alder

Set vi inn i formelen ovanfor finn vi at forventa sannsyn p(alder) =

1/(1+exp[-(-4.95993103+ 0.1555908*Alder –0.00141456*Alder*Alder)])

d) Kva er definisjonen av Oddsen for å vitje lokalt kunstgalleri for den persontypen som er definert i pkt c)? Bruk definisjonen og modell 3 til å finne oddsraten for å velje å vitje lokalt kunstgalleri mellom ein mann med 19 års utdanning og ein med 18 års utdanning.

Odds
Oddsen er definert som sannsynet for å vitje lokalt kunstgalleri dividert med ein minus sannsynet for å vitje lokalt kunstgalleri. Logiten er definert som den naturlege logaritmen til oddsen. Dermed vil vi finne oddsen ved å opphøge grunntalet e i Logiten; dvs. $O_i = \exp\{L(i)\}$, der i= person av typen definert i pkt c.

Oddsraten
Oddsraten finn vi da som høvestalet mellom to Odds. La j= person av typen "i" men med 18 års utdanning i staden for 19. Da er Oddsraten (i-person i høve til j-person) = $O_i / O_j = \exp[L(i)] / \exp[L(j)] = \exp[L(i)-L(j)]$

Det kan her argumenterast med at modellen er estimert med 17 års utdanning som høgaste verdi og der alle med 18 og 19 års utdaninng er koda 17. Ved å nytte samme koding som i datamaterialet vil begge får alder 17 og oddsraten må da bli 1. Dette svaret må aksepterast.

Gitt at vi ønskjer å nytte den estimerte modellen til å ekstrapolere til litt høgare utdaningar vil vi i modell 3 finne at
L(i) = -2.2630382 +0.26434937 Kvinne -0.2060165 E.utdanning +0.13107263 Mors utdanning -0.1325838 Alder +0.01652043 E.utdanning*Alder +0.00135636 Alder*Alder -0.0001597 Alder*Alder*E.utdanning
Sidan den einaste skilnaden mellom i-personar og j-personar ser at j-personane har verdien E.utdanning-1 i staden for E.utdanning får vi at
L(j) = -2.2630382 +0.26434937*Kvinne -0.2060165*(E.utdanning –1) +0.13107263*Mors utdanning -0.1325838*Alder +0.01652043*(E.utdanning -1) *Alder +0.00135636*Alder*Alder -0.0001597*Alder*Alder*(E.utdanning –1)

Dermed blir differansen L(i)-L(j) =
-0.2060165 E.utdanning – (-0.2060165 (E.utdanning –1))
+0.01652043 E.utdanning*Alder - 0.01652043 (E.utdanning -1)*Alder
-0.0001597 Alder*Alder*E.utdanning – (-0.0001597Alder*Alder*(E.utdanning –1)) =
-0.2060165 +0.01652043*Alder -0.0001597 Alder*Alder

Set vi inn verdiane frå punkt c) finn vi no at
L(i)-L(j) = -0.2060165 +0.01652043*50 -0.0001597 *50*50 = 0.220755
Oddsraten for å velje å vitje lokalt kunstgalleri mellom ein mann med 19 års utdanning og ein med 18 års utdanning blir da **exp[0.220755] = 1.247**