EKSAMENSOPPGÅVER SVSOS316
HAUST 2001
FRAMLEGG TIL LØYSING

Erling Berge
Institutt for sosiologi og statsvitskap
Norges Teknisk Naturvitskapelege Universitet

«Bruksanvisning»

Når ein går i gang 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 i eit slikt høve er tale om. 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 fordelinga av 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 alle oppgåvene. I arbeidet med dette løysingsframlegget har det vore gjort meir arbeid enn det 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 lesing 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 kort korleis ein definerer DFBETAS for case nr i og forklaringsvariabel nr k.

Dersom vi først estimer regresjonskoeffesienten b_k på heile utvalet og etterpå estimerer same koeffesienten på eit utval der case nr i er utelaten, vil vi finne verdien $b_{k(i)}$. Vi kan sjå på differansen $b_k - b_{k(i)}$ som uttrykk for kor stor innverknad case nr i har på regresjonsresultatet. For å vurdere om denne er stor eller liten standardiserer vi differansen ut frå standardfeilen i regresjonen utan case nr i, dvs.: $s_{e(i)} / \sqrt{RSS_k}$ der $s_{e(i)}$ er standardavviket til residualen i regresjonen utan case i og RSS_k er "Residual Sum of Squares" i regresjonen av x_k på resten av x-variablane i modellen der case nr i er med.

DFBETAS er da definert som $DFBETAS_{ik} = [b_k - b_{k(i)}] / [s_{e(i)} / \sqrt{RSS_k}]$

b) Forklar kort framgangsmåten for estimering av Tobitmodellar på sensurerte data.

Dersom vi i eit gitt datamateriale manglar opplysningar om Y for gitte verdiar av Y medan vi har opplysningar om x-variablane, er utvalet sensurert. Vi kan estimere den ordinære lineærmodellen

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ ... + \beta_{K-1} X_{K-1\,i} + \varepsilon_i \ , \ i = 1,2,3, \ \ldots \ ,n$$

berre for dei einingane der y er observert.

Tobitmodellar har som føresetnad at vi kjenner verdien av Y for $Y > 0$ men ikkje dersom $Y \leq 0$. Når Y eigentleg er mindre enn 0 vert Y automatisk sett til 0. I dette høvet kan det visast at forventninga til Y i modellen ovanfor vert

$$E[Y_i \mid X_{1i} , ... , X_{K-1\,i}] = Pr\{ Y_i > 0 \mid X_{1i} , ... , X_{K-1\,i} \} E[Y_i \mid Y_i > 0, X_{1i} , ... , X_{K-1\,i}]$$
$$+ Pr\{ Y_i \leq 0 \mid X_{1i} , ... , X_{K-1\,i} \} E[Y_i \mid Y_i \leq 0, X_{1i} , ... , X_{K-1\,i}]$$

Men sidan Y_i alltid vert sett til 0 når $Y_i \leq 0$ vil $E[Y_i \mid Y_i \leq 0, X_{1i} , ... , X_{K-1\,i}] = 0$. Dermed vert

$$E[Y_i \mid X_{1i} , ... , X_{K-1\,i}] = Pr\{ Y_i > 0 \mid X_{1i} , ... , X_{K-1\,i} \} E[Y_i \mid Y_i > 0, X_{1i} , ... , X_{K-1\,i}]$$

Vi ser at i eit sensurert utval med sensurering ved 0 vil regresjonsmodellen eigentleg vere produktet av to delmodellar. I den eine delmodellen, $Pr\{ Y_i > 0 \mid X_{1i} , ... , X_{K-1\,i} \}$, må vi estimere sannsynet for å ha ein verdi på den avhengige variabelen gitt dei observerte x-ane. I den andre delmodellen estimerer vi forventa Y gitt x-ane og det faktum at personen heilt sikkert har ein verdi over 0 på Y-variabelen.

Sidan vi har observert x-ane også når vi manglar observerte Y-verdiar, kan vi lage ein dummyvariabel for manglande Y-verdi og nytte ein probitmodell til å estimere sannsynet for å ha Y-verdi ved hjelp av dei observerte x-ane. Deretter estimerer vi den andre delmodellen ved hjelp av kunnskapen frå probitmodellen. Den endelege modellen blir til som produktet av dei to delmodellane.

Den andre delmodellen kan estimerast ved hjelp av OLS-metoden, men den vil innehalde ein ekstra parameter på grunn av vilkåret $Y_i > 0$. Dette fører til at forventninga til feilleddet i modellen vert større enn 0. Det kan visast at vi i

staden for det vanlege feilleddet vil ha eit ledd som kan skrivast $\sigma/(\phi_i/\Phi_i)$. Høvetalet (ϕ_i/Φ_i) er høvet mellom sannsynet for Y=1 gitt $\mathbf{X} = (X_{1i}, \ldots, X_{K-1\,i})$ og sannsynet for at $Y \neq 1$ gitt $\mathbf{X} < (X_{1i}, \ldots, X_{K-1\,i})$. Dette høvetalet vert kalla hasardraten og vert oftast symbolisert ved λ_i. Generelt kan ein seie at ei hasardrate er sjansen for suksess (Y=1) dividert med sjansen for ikkje å ha hatt suksess fram til det punkt ein vert sett å ha det.

Den andre delmodellen vil da sjå slik ut:
$E[Y_i \mid Y_i > 0, X_{1i}, \ldots, X_{K-1\,i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{K-1} X_{K-1i} + \sigma \lambda_i$
der vi finn λ_i ved hjelp av den først estimerte probitmodellen.

Denne måten å estimere Tobitmodellar på vert kalla Heckman sin to-stegsprosedyre. Den gir likevel gale estimat av strandardfeilane til perametrane og av σ . Ved omformulering av modellen til ein integrert modell og bruk av maksimum likelihood teknikk kjem rund dei problema.

OPPGÅVE 2 (OLS-regresjon, vekt 0,45)
I tabellvedlegget til oppgåve 2 er det estimert 16 modellar av eiga inntekt (E.inntekt).

a. **Ta utgangspunkt i modell 1 og forklar med ord kva effekten av E.utdanning tyder. Finn så i same modellen eit konfidensintervall for effekten av E.utdanning. Finn ut frå modell 4 forventa inntekt for ei 40 år gammal kvinne med 12 års utdanning og heiltidsarbeid i Kværner Oil&Gas. Test om det er ein lineær eller kurvelineær samanheng mellom alder og inntekt.**

Ta utgangspunkt i modell 1 og forklar med ord kva effekten av E.utdanning tyder.
Den avhengige variabelen E.inntekt er målt i 1000 kroner. Variabelen E.utdanning er målt i år. Estimatet av modell 1 viser at for kvart år ekstra utdanning ein person får, vil inntekta auke med 5,25*kr 1000 = kr 5.250 om alt anna er likt.

Finn så i same modellen eit konfidensintervall for effekten av E.utdanning.
Det er knytt uvisse til talet som gir oss inntektseffekten av kvart år utdanning. Kor stor uvissa er kan vi få vite noko om ved å finne eit konfidensintervall for effekten.

I modell 1 er $b_{E.utdanning}$, effekten av E.utdanning, oppgitt til å vere 5,2498587 med ein standardfeil på 0,434555. I estimatet er det nytta eit datamateriale på n=2634 observasjonar av E.inntekt. Modellen har K=5 parametrar slik at t-testane vil ha n-K = 2629 fridomsgrader.

Dersom vi kan gå ut frå at feilledda er normalfordelte vil eit 95% konfidensintervall (5% signifikansnivå) vere gitt ved

$$b_{E.utdanning} - SE_{E.utdanning} * t_{5\%} < \beta_{\mathbf{E.utdanning}} < b_{E.utdanning} + SE_{E.utdanning} * t_{5\%}$$

der b er den estimerte regresjonskoeffesienten, SE er standardfeilen til regresjonskoeffesienten og $t_{5\%}$ er fraktilen i t-fordelinga i ein tosidig test med signifikansnivå 0,05. I følgje tabell A4.1 hos Hamilton (1992:350) vil vi med meir enn 120 fridomsgrader (vi har 2629 fridomsgrader) ha at $t_{5\%}$ =1,96 (tosidig test; 1,96 = $t_{2,5\%}$ i einsidig test). Set vi inn i formelen finn vi at

$$5,2498587 - 0,434555 * 1,96 < \beta_{E.utdanning} < 5,2498587 + 0,434555 * 1,96$$
$$5,2498587 - 0,8517278 < \beta_{E.utdanning} < 5,2498587 + 0,8517278$$

$4{,}3981309 < \beta_{\mathbf{E.utdanning}} < 6{,}1015865$

Eller sagt med ord: i 95% av alle tilfeldige utval frå den same populasjonen vil vi finne at den verkelege effekten av eitt års ekstra utdanning ligg mellom 4.398 og 6.101 kroner om alt anna er likt.

Finn ut frå modell 4 forventa inntekt for ei 40 år gammal kvinne med 12 års utdanning og heiltidsarbeid i Kværner Oil&Gas.
Vi finn forventa inntekt for ein slik person ved å setje variabelverdiane inn i likninga i modell 4. Ein finn at forventa inntekt er omlag 192.000 kroner.

Variabel	Variabelverdi	Multiplisert med	Parameter Estimat	Gir resultatet
Konstant			-87,86921	-87,86921
Alder	40	*	4,8479342	193,91737
Alder*Alder	40*40	*	-0,045279	-72,44675
Mann	0	*	-71,52876	0
E,utdanning	12	*	6,9131003	82,95720
Heiltidsarbeid	1	*	75,334491	75,33449
Offentleg sektor	0	*	-6,776133	0
Alder*Mann	40*0	*	5,0725196	0
Alder*Alder*Mann	40*40*0	*	-0,04956	0
Forventa verdi av E,inntekt				191,89310

Test om det er ein lineær eller kurvelineær samanheng mellom alder og inntekt.
I modell 1 er ikkje alder inkludert som variabel, i modell 2 er alder inkludert lineært og i modell 3 er alder inkludert som kurvelinær med eit andregradspolynom. I modell 2 gir den lineære variabelen alder eit klart signifikant bidrag til å forklare variasjon i inntekt. Dersom nullhypotesa $\beta_{\mathbf{Alder}} = 0$ er rett, vil t-observatoren i modell 2 ha verdien 14.22. Dette er eit svært lite sannsynleg resultat. Vi forkastar derfor nullhypotesa. Vi kan akseptere alder som lineær.

Vi kan teste modell 3 mot modell 1 med ein F-test. Når vi samanliknar to modellar estimert på same utval av n case, ein modell med K parametrar 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. Set vi inn tal frå modellane 1 og 3 finn vi

$$F^{2}_{2627} = \frac{(RSS[7\text{-}2] - RSS[7]) / 2}{RSS[7] / (2634\text{-}7)}$$

$$= \frac{(11049201 - 9334027) / 2}{9334027 / 2627}$$

$$= 241.362174$$

I F-fordelinga med 2 og uendeleg mange fridomsgrader er sannsynet for å få ein F-verdi større enn 3.00 mindre enn 0,05. Vi må klart forkaste 0-hypotesa at alderspolynomet ikkje har effekt i modellen.

Spørsmålet blir om alder som kurvelineært variabel er betre enn alder som lineær variabel. Dette er ikkje opplagt ut frå dei to testane vi har gjort så langt, og ein F-test av kvadratleddet aleine gir ikkje meir kunnskap enn det t-testen av koeffisienten for kvadratleddet gir. F-testen gir ein F-verdi på 260.623877. Dette er det same som t-verdien kvadrert (T-verdien kvadrert gir 260.4996. Skilnaden skuldast avrundingsfeil. T-observatoren er gitt med 2 desimalar.) Vi må studere nærmare kva som skjer når vi går frå modell 2 til modell 3.

Dersom vi testar nullhypotesane $\beta_{\mathbf{Alder}} = 0$ og $\beta_{\mathbf{Alder*Alder}} = 0$ i modell 3 får vi t-verdiane $t_{\mathbf{Alder}} = 18.60$ og $t_{\mathbf{Alder*Alder}} = -16.14$. Begge testane fører til at vi forkastar nullhypotesen, og begge t-verdiane er større enn t-verdien for testen av nullhypotesen $\beta_{\mathbf{Alder}} = 0$ i modell 2. Dette resultatet finn vi trass i at multikollineariteten mellom Alder og Alder*Alder vil gi større standardavvik for det einskilde parameterestimatet i modell 3. Vi legg også merke til at determinasjonskoeffesienten aukar frå 0,484 i modell 2 til 0,531 i modell 3.

Konklusjonen må vere at alder som kurvelineær variabel er klart betre enn alder som lineær variabel.

Generelt kan ein seie at dersom begge parametrane i eit alderspolynom er signifikant ulik 0, vil heile polynomet yte signifikant til forklaringa av variasjonen i den avhengige variabelen. Meir omfattande testing trengst berre dersom ein av parametrane ikkje er signifikant ulik 0.

b. Formuler den modellen som er estimert som Modell 8. Vurder om testane i modell 8 er truverdige. Drøft mogelege forbetringar av modellspesifikasjonen.

Formuler den modellen som er estimert som Modell 8.
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 8 er følgjande variablar definert:

Variabel		**Variabelnamn**	**Kommentar**
Y	=	E.inntekt	Inntekt i 1000 kroner
X_1	=	Alder	Alder er inkludert kurvelineært ved eit polynom
X_2	=	Alder*Alder	
X_3	=	Mann	Dummykoda
X_4	=	E.utdanning	Utdanning i år
X_5	=	Heiltidsarbeid	Dummykoda
X_6	=	Offentleg sektor	Dummykoda
X_7	=	Alder*Mann	Interaksjonsledd mellom kjønn og Alder
X_8	=	Alder*Alder*Mann	
X_9	=	E.utdanning*Mann	Interaksjonsledd
X_{10}	=	Heiltidsarbeid*Mann	Interaksjonsledd
X_{11}	=	Alder*Heiltidsarbeid	Interaksjonsledd mellom Heiltidsarbeid og Alder
X_{12}	=	Alder*Alder*Heiltidsarbeid	
X_{13}	=	E.utdanning*Heiltidsarbeid	Interaksjonsledd
X_{14}	=	Offentleg sektor*Heiltidsarbeid	Interaksjonsledd

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.

I populasjonen føreset 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

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

når vi lar i gå over heile populasjonen. Lar vi k=0, 1, 2, ... ,14, 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

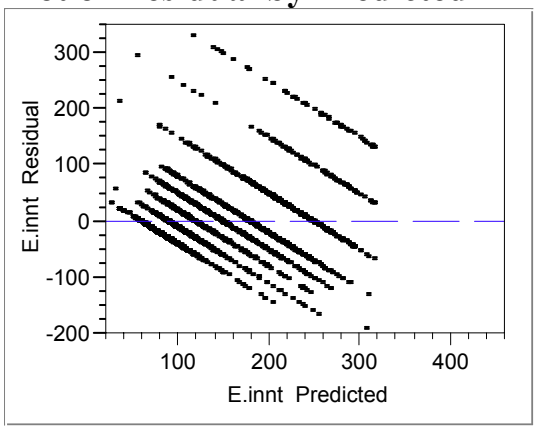
II. Gauss-Markov krava for «Best Linear Unbiased Estimates» (BLUE) er oppfylt, dvs.:
- Faste x-verdiar (dvs. vi kan i prinsippet trekke nye utval med same x-verdiar men der vi får ulike y-verdi på grunn av den stokastiske komponenten i feilleddet).
- Feilledda har forventning 0 for alle i, dvs: $E(\varepsilon_i) = 0$ for alle i .
- Feilledda har konstant varians (homoskedastisitet) dvs: $var(\varepsilon_i) = \sigma^2$ for alle i .
- Feilledda er ukorrelerte med kvarandre (det er ikkje autokorrelasjon) dvs: $cov(\varepsilon_i, \varepsilon_j) = 0$ for alle $i \neq j$.

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

Vurder om testane i modell 8 er truverdige.
Generelt veit vi at tekniske og substansielle problem som t.d. ikkje-lineære samband, utelatne variablar, målefeil i dei uavhengige variablane, heteroskedastisitet, autokorrelasjon og ikkje-normalfordelte feilledd vil føre til at t- og F-testane ikkje blir truverdige. Autokorrelasjon og heteroskedastisitet gir skeive estimat av variansane og fører til upålitelege testar. Dersom residualen ikkje er normalfordelt, er ikkje test-observatorane (F- og t-observatorane) definert, og kan sjølvsagt ikkje nyttast.

Autokorrelasjon er ikkje aktuelt i eit sannsynsutval av personar. Av dei andre faktorane som kan påverke truverdet til testane er det berre heteroskedastisitet og ikkje-normalfordelte residualar som kan undersøkjast ved hjelp av vedlagte tabellar.

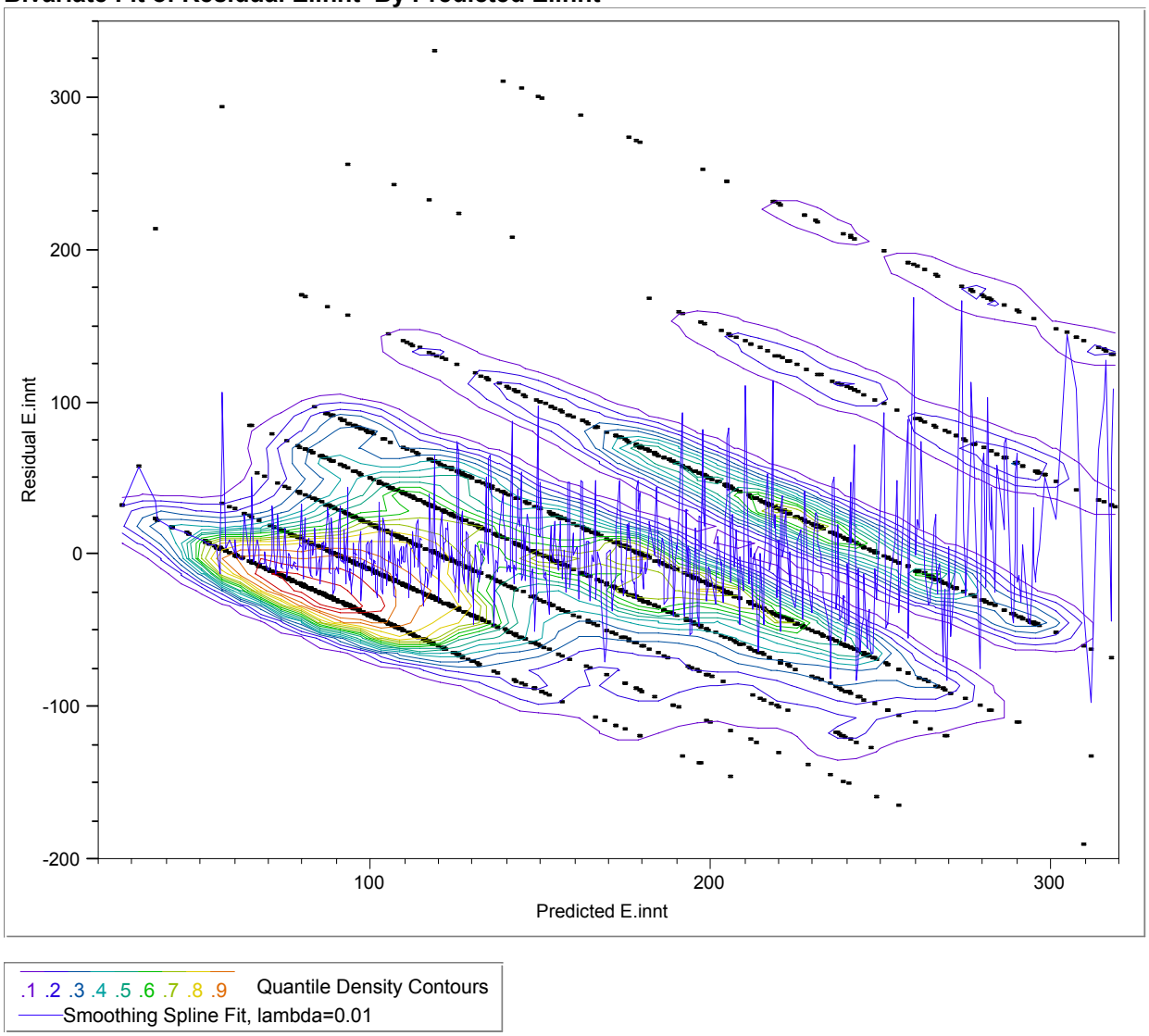
Plot of Residual by Predicted Y in Model 8



Ser vi først på plottet av residualen mot predikert verdi av Y i modell 8 ser vi dei systematiske linjene som følgjer av få verdiar (8) på avhengig variabel. Linjene som sluttar øvst til venstre i diagrammet ser ut til å ha færre observasjonar enn resten av linjestykka. Spreiinga av residualane for dei lågaste verdiane av predikert Y vert då mindre enn for dei som er litt større der vi får fleire relativt større residualar, særleg på negativ side. Diagrammet tyder altså på ein viss heteroskedastisitet. Det kan likevel ofte vere vanskeleg å avgjere om ein har heteroskedastistet berre på grunnlag av ein slik figur.

Ekstra: dette er utanom eksamen:

Bivariate Fit of Residual E.innt By Predicted E.innt

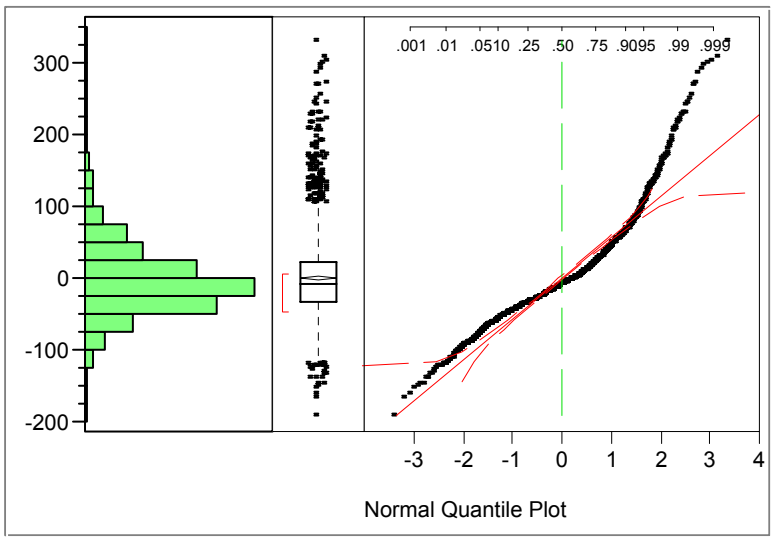


Vi kan gå attende til data og sjå nærmare på korleis residualane fordeler seg. Ved hjelp av tettleiksellipsar kan vi sjå meir systematisk på kor mange case som har same verdien på residualen og predikert Y, og som dermed blir plotta inn i same punkt i diagrammet (tettleiken til fordelinga av residual mot predikert verdi). Vi kan også legge inn ein splinefunksjon med mange bitar for å sjå korleis tyngdepunktet av variabelverdiane varierer lokalt. Det ser også ut frå dette ut til at det er ein viss tendens til aukande variasjon i residualen når predikert Y veks. Vi har med andre ord ein svak heteroskedastisitet.

Slutt ekstrastoff

Ser vi på histogrammet av residualen ser vi at fordelinga er høgreskeive (relativt mange fleir store positive residualar enn store negative). I samsvar med dette finn vi at medianen er lik –8.4, dvs. mindre enn gjennomsnittet (som sjølvsagt er 0).
Likeeins finn vi at IQR/1.35 = 41.41 som er mindre enn standardavviket på 57.17. Den observerte fordelinga av residualane er ikkje berre høgreskeiv, den har også tyngre halar (fleire utliggarar) enn ei tilsvarande normalfordeling.

Det same ser vi i kvantil-normal plottet. Avviket frå normalfordeling er så pass tydeleg at det kan vere grunn til å tvile på testane i modell 8.

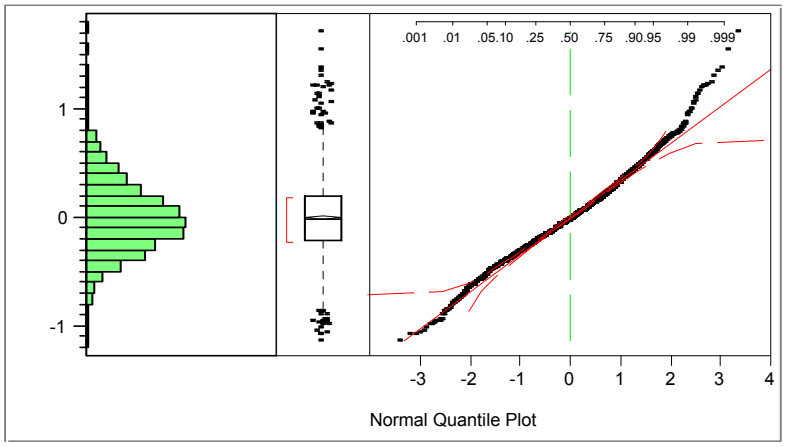


Samla sett vil ein vel seie at tendensen til heteroskedastisitet saman med avvika frå normalfordelinga (dette kan godt vere to sider av same fenomen) tilseier stor varsemd med å tru på testane i modellen.

Drøft mogelege forbetringar av modellspesifikasjonen.
Ein betre modellspesifikasjon gjennom fleire variable eller gjennom transformasjon av avhengig variabel kan vere med å gi større tiltru til konklusjonane. I modellane 12 og 14 vert variablar som "Kjelde til Livsopphald" og "Bustad" lagt til og i modell 16 er det gjort ein transformasjon av Y. Modellane 12 og 14 inneheld ikkje nok opplysningar til å kunne vurdere om dei har mindre grad av heteroskedastisitet og om dei har residualar som i større grad nærmar seg normalfordelinga.

I modell 16 er den avhengige variabelen transformert. Dette ser ut til å ha ført til ei betre fordeling av residualane. Dei er mindre skeivfordelt.

Figur: Kvantil-normal plott for residualane i modell 16



Plottet av residualen mot predikert verdi ser ut til å ha endra seg lite. Men om det var vanskeleg å avgjere grad av heteroskedastisitet i utgangspunktet er det enno verre å vurdere om modell 16 er ei forbetring av situasjonen. Det er likevel lite truleg at graden av heteroskedastisitet er vorten verre. Truleg vil modell 16 vere ein betre spesifikasjon enn modell 8.

c. Test om Kjelde til livsopphald gir ei signifikant yting til å forklare variasjonen i inntekt. Finn ved hjelp av modell 12 eit estimat av skilnaden i inntekt mellom menn og kvinner når ein føreset at dei er funksjonærar i heiltidsarbeid, har 15 års utdanning og arbeider i Direktoratet for naturforvaltning. Ta utgangspunkt i modell 16 og skriv ut formelen for eit betinga effekt plott for samanhengen mellom alder og inntekt for personar med heiltidsarbeid, 15 års utdanning og arbeid i offentleg sektor.

Test om Kjelde til livsopphald gir ei signifikant yting til å forklare variasjonen i inntekt.
Når vi skal teste om **"Kjelde til livsopphald"** yter signifikant til å forklare variasjonen i **E.inntekt**. må vi gå ut frå at føresetnadene som er presisert ovanfor er rette.

Kjelde til livsopphald er dummykoda. Kategorien «**Arbeider**» er utelaten og fungerer som referansekategori. Dei seks andre kategoriane i variabelen er inkludert i modell 11, 12 og 14. Når ein fjernar dei inkluderte kategoriane av **Kjelde til livsopphald** frå modell 11 får vi modell 10, fjernar vi dei frå modell 12 får vi modell 8 og fjernar vi dei frå modell 14 får vi modell 13. Det er med andre ord 3 måtar å konstruere ein F-test for å avgjere om **Kjelde til livsopphald** yter signifikant til å forklare variasjonen i **E.inntekt**. Kva for ein test er best? For å avgjere det må vi sjå nærmare på dei tre modellane 8, 10 og 13.

Modellane 9 og 10 viser at interaksjonsledda mellom **Offentleg sektor** og **Alder, Ma**nn og **E.utdanning** ikkje er signifikante kvar for seg og samanlikning av modell 10 med modell 8 viser at determinasjonskoeffesienten aukar med berre 0,001076. Vi kan nytte ein F-test for å avgjere om dette er ein signifikant auke.

Når vi samanliknar to modellar estimert på same utval av n case, ein modell med K parametrar 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.

I modell 8 og 10 finn vi

Analysis of Variance Model 8

Source	DF	Sum of Squares	Mean Square	F Ratio
Regression	14	11291092	806507	245.4795
Residual	2619	8604550	3285	Prob > F
C. Total	2633	19895642		0.0000

Analysis of Variance Model 10

Source	DF	Sum of Squares	Mean Square	F Ratio
Regression	18	11312504	628472	191.4749
Residual	2615	8583138	3282	Prob > F
C. Total	2633	19895642		0.0000

slik at

$$F^{4}_{2634-19} = \frac{(RSS[19\text{-}4] - RSS[19]) / 4}{RSS[19] / (2634 - 19)}$$

$$= \frac{(8604550 - 8583138) / 4}{8583138/2615} = 1.63$$

Vi kan da nytte testobservatoren F på modellane 10 (stor modell) og 8 (liten modell) for å samanlikne to «nesta» modellar. Sidan 5 % fraktilen(α) i F-fordelinga med 4 og 2615 fridomsgrader er 2.37 kan vi ikkje forkaste nullhypotesa om at dei fire variablane ikkje har nokon effekt. Dei fire interaksjonsledda er med andre ord irrelevante variablar og bør ikkje vere med i dei vidare drøftingane.

Å teste **Kjelde til livsopphald** ved å samanlikne modell 10 og 11 gir derfor ikkje den beste testen. Alternativa, å samanlikne modell 8 og 12 eller modell 13 og 14, er betre. Vi skal likevel gjennomføre testen ved samanlikning av modell 11 mot 10.

For modell 11 har vi gitt

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Regression	24	11825810	492742	159.3049
Residual	2609	8069832	3093	Prob > F
C. Total	2633	19895642		0.0000

Testen av modell 11 mot modell 10 gir
H = 6
K = 25
n-K= 2634 - 25 = 2609
RSS[K-H] = 8583138
RSS[K] = 8069832

Dette gir $F^{6}_{2609} = 27.6588855$, og sidan 5 % fraktilen(α) i F-fordelinga med 6 og 2609 fridomsgrader er 2.10 (Hamilton 1992, tabell A4.2) vil vi forkaste nullhypotesa om at variabelen «Kjelde til livsopphald» ikkje yter til å forklare variasjonen i eiga inntekt.

I den alternative samanlikninga av modell 12 mot modell 8 har vi gitt

Analysis of Variance Model 12

Source	DF	Sum of Squares	Mean Square	F Ratio
Regression	20	11808802	590440	190.7816
Residual	2613	8086840	3095	Prob > F
C. Total	2633	19895642		0.0000

Dette gir
H = 6,
K = 21,
n-K= 2634 - 21 = 2613,
RSS[K-H] = 8604550,
RSS[K] = 8086840,
og $F^{6}_{2613} = 27.8801986$. Sidan 5 % fraktilen($\alpha$) i F-fordelinga med 6 og 2613 fridomsgrader er 2.10 (Hamilton 1992, tabell A4.2) vil vi forkaste nullhypotesa om at variabelen «Kjelde til livsopphald» ikkje yter til å forklare variasjonen i eiga inntekt.

Skilnaden mellom samanlikningane 8-12 og 13-14 er variabelen **"Bustad"**. Den er ny i modellane 13 og 14. Dersom bustad er ein relevant variabel for modellen gir det ein betre test av "**Kjelde til livsopphald**" å inkluderer bustad i testen. For å avgjere om bustad er ein relevant variabel må vi samanlikne modellane 8 og 13.

For modell 13 har vi gitt

Analysis of Variance Model 13

Source	DF	Sum of Squares	Mean Square	F Ratio
Regression	19	11391594	599558	184.2938
Residual	2614	8504048	3253	Prob > F
C. Total	2633	19895642		0.0000

Test "Bustad" ved samanlikning av modell 13 mot modell 8 gir
H = 5
K = 20
n-K= 2634 - 20 = 2614
RSS[K-H] = 8604550
RSS[K] = 8504048

Dette gir $F^5_{2614} = 6.17852176$, og sidan 5 % fraktilen(α) i F-fordelinga med 5 og 2614 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 sterkaste testen av **Kjelde til livsopphald** vil vere å teste den ved å samanlikne modellane 13 og 14 heller enn 8 og 12 sidan ein da tar omsyn også til effekten av bustad før ein vurderer effekten av **Kjelde til livsopphald**.

Ut frå variansanalysetabellen frå modell 14 (nedanfor) og 13 (ovanfor)

Analysis of Variance Model 14

Source	DF	Sum of Squares	Mean Square	F Ratio
Regression	25	11896297	475852	155.1404
Residual	2608	7999345	3067	Prob > F
C. Total	2633	19895642		0.0000

Finn vi
H = 6
K = 26
n-K= 2634 - 26 = 2608
RSS[K-H] = 8504048
RSS[K] = 7999345

Dette gir $F^6_{2608} = 27.4244417$, og sidan 5 % fraktilen(α) i F-fordelinga med 6 og 2608 fridomsgrader er 2.10 (Hamilton 1992, tabell A4.2) vil vi forkaste nullhypotesa om at variabelen «**Kjelde til livsopphald**» ikkje yter noko til å forklare variasjonen i eiga inntekt.

Finn ved hjelp av modell 12 eit estimat av skilnaden i inntekt mellom menn og kvinner når ein føreset at dei er funksjonærar i heiltidsarbeid, har 15 års utdanning og arbeider i Direktoratet for naturforvaltning.

Sidan 15 års utdanning ikkje er ein verdi som er i bruk på variabelen E.utdanning er det akseptabelt å nytte verdien 14 (begge verdiane er for så vidt problematiske). Legg merke til at alder ikkje er nemnt. Vi må da bruke vilkårleg

alder. Vi kallar verdien ”Alder”. Vi finn skilnaden mellom menn og kvinner ved først å sette menn sine verdiar inn i likninga for modell 12 og sidan kvinner sine verdiar og så ta differansen av dei to likningane.

Modell 12 av E.inntekt:

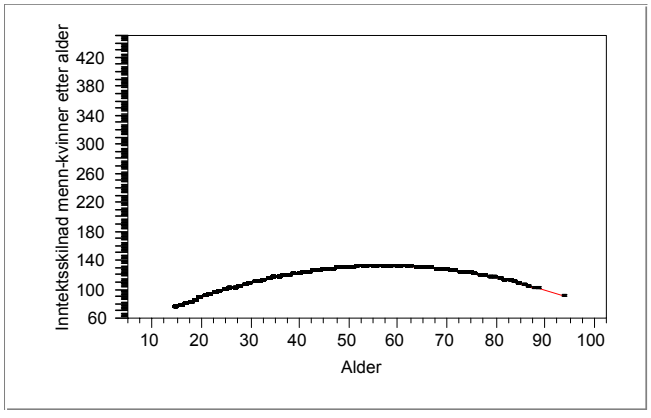
Variabelnamn	Variabelverdi Menn	Multiplisert med Parameterestimat	-	Variabelverdi Kvinner	Multiplisert med Parameterestimat	= Skilnaden i inntekt mellom menn og kvinner
Konstant		19.96233			19.96233	0
Alder	Alder	1.8827734		Alder	1.8827734	0
Alder*Alder	Alder*Alder	-0.017426		Alder*Alder	-0.017426	0
Mann	1	-70.71269		0	-70.71269	-70.71269
E.utdanning	15	3.2568789		15	3.2568789	0
Heiltidsarbeid	1	-106.4836		1	-106.4836	0
Offentleg sektor	1	10.92664		1	10.92664	0
Alder*Mann	Alder*1	3.6163248		Alder*0	3.6163248	Alder*3.6163248
Alder*Alder*Mann	Alder*Alder*1	-0.031424		Alder*Alder*0	-0.031424	Alder*Alder*(-0.031424)
E.utdanning*Mann	15*1	1.3197444		15*0	1.3197444	15*1.3197444
Heiltidsarbeid*Mann	1*1	22.48204		1*0	22.48204	22.48204
Alder*Heiltidsarbeid	Alder*1	6.3899222		Alder*1	6.3899222	0
Alder*Alder*Heiltidsarbeid	Alder*Alder*1	-0.074908		Alder*Alder*1	-0.074908	0
E.utdanning*Heiltidsarbeid	15*1	3.8438286		15*1	3.8438286	0
Offentleg sektor*Heiltidsarbeid	1*1	-31.1802		1*1	-31.1802	0
Funksjonær	1	31.591156		1	31.591156	0
Sjølvstendig	0	20.759837		0	20.759837	0
Elev/ student	0	-25.6586		0	-25.6586	0
Pensjon/ trygd	0	-10.51641		0	-10.51641	0
Andre KtL	0	-6.515989		0	-6.515989	0
Uoppgitt KtL	0	6.945728		0	6.945728	0

Summerer vi kollonna til høgre finn vi at differansen mellom menn og kvinner vert lik

-70.71269 + Alder*3.6163248 + Alder*Alder*(-0.031424) + 19.796166 + 22.48204 =

-28.434484 + Alder*3.6163248 + Alder*Alder*(-0.031424) .

Differansen varierer altså med alderen slik diagrammet nedanfor illustrerer:



Ta utgangspunkt i modell 16 og skriv ut formelen for eit betinga effekt plott for samanhengen mellom alder og inntekt for personar med heiltidsarbeid, 15 års utdanning og arbeid i offentleg sektor.

Vi ser at vi ikkje har oppgitt kjønn. Vi vil da halde på "Mann" som variabel i formelen for eit betinga effekt plott. Vi kan da seinare sette inn verdien 1 for menn og 0 for kvinner:

Modell 16 av Ln(E.innt)

Variabelnamn	Variabelverdi	Parameterestimate	=	Parameterestimate* Variabelverdi
Konstant		3.1905927		3.1905927
Alder	Alder	0.0418769		0.0418769*Alder
Alder*Alder	Alder*Alder	-0.000399		-0.000399* Alder*Alder
Mann	Mann	-0.223609		-0.223609* Mann
E.utdanning	15	0.0308258		0.0308258*15
Heiltidsarbeid	1	0.0965473		0.0965473*1
Offentleg sektor	1	0.1590762		0.1590762*1
Alder*Mann	Alder*Mann	0.0151868		0.0151868*Alder*Mann
Alder*Alder*Mann	Alder*Alder*Mann	-0.0001		-0.0001*Alder*Alder*Mann
E.utdanning*Mann	15*Mann	-0.001286		-0.001286*15*Mann
Heiltidsarbeid*Mann	1*Mann	0.0781833		0.0781833*1*Mann
Alder*Heiltidsarbeid	Alder*1	0.0247525		0.0247525*Alder*1
Alder*Alder*Heiltidsarbeid	Alder*Alder*1	-0.000369		-0.000369*Alder*Alder*1
E.utdanning*Heiltidsarbeid	15*1	0.0126049		0.0126049*15*1
Offentleg sektor*Heiltidsarbeid	1*1	-0.24989		-0.24989*1*1

Dersom vi no summerer kollonna til høgre og stokkar om på rekkjefølgja av ledda i summen finn vi

Ln(E.inntekt) =
3.1905927 + 0.0965473 + 0.1590762 - 0.24989 + 0.0308258*15 + 0.0126049*15
+ 0.0418769*Alder + 0.0247525*Alder - 0.000399* Alder*Alder - 0.000369*Alder*Alder
- 0.223609* Mann+ 0.0781833*Mann - 0.001286*15*Mann
+ 0.0151868*Alder*Mann - 0.0001*Alder*Alder*Mann

Dette er det same som at
Ln(E.inntekt) =
3.8477867
+ (0.0418769 + 0.0247525)*Alder - (0.000399 + 0.000369)*Alder*Alder
+ (- 0.223609 + 0.0781833 - 0.001286*15)*Mann
+ (0.0151868*Alder - 0.0001*Alder*Alder)*Mann

Dette gir vidare at
Ln(E.inntekt) =
3.8477867
+ 0.0666294*Alder - 0.000768*Alder*Alder
- 0.1647157*Mann
+ (0.0151868*Alder - 0.0001*Alder*Alder)*Mann =
Ln(E.inntekt) =
3.8477867 + 0.0666294*Alder - 0.000768*Alder*Alder
- 0.1647157*Mann + (0.0151868*Alder - 0.0001*Alder*Alder)*Mann =
Ln(E.inntekt) =
3.8477867 + 0.0666294*Alder - 0.000768*Alder*Alder +
(- 0.1647157 + 0.0151868*Alder - 0.000100*Alder*Alder)*Mann

Tar vi no den inverse transformasjonen av E.inntekt finn vi at

**E.inntekt = exp{3.8477867 + 0.0666294*Alder - 0.000768*Alder*Alder +
(- 0.1647157 + 0.0151868*Alder - 0.000100*Alder*Alder)*Mann }**

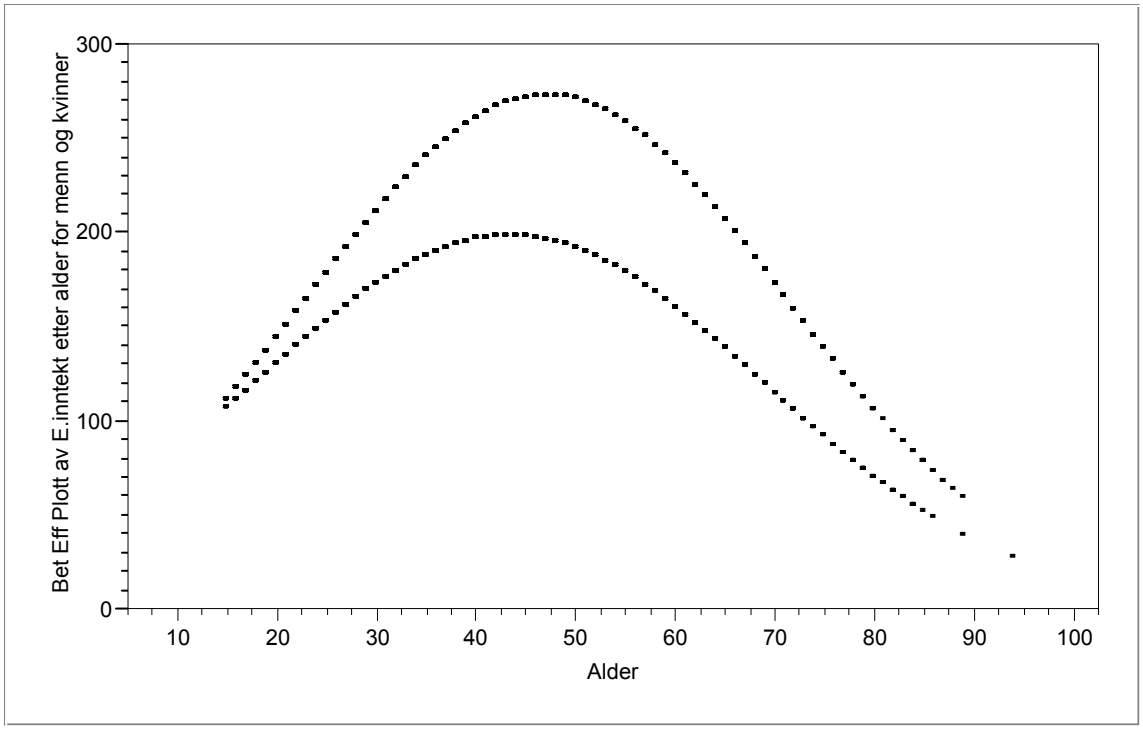
For kvinner får vi da at inntekta varierer med alderen etter samanhengen:
**$E.inntekt_{Kvinner}$ = exp{3.8477867 + 0.0666294*Alder -
0.000768*Alder*Alder}**

Og tilsvarande for menn:
$E.inntekt_{Menn}$ =
exp{3.8477867 + 0.0666294*Alder - 0.000768*Alder*Alder +
(- 0.1647157 + 0.0151868*Alder - 0.000100*Alder*Alder)*1 } =
exp{3.8477867 - 0.1647157 + (0.0666294 + 0.0151868)*Alder – (0.000768 +
0.000100) *Alder*Alder} =
exp{3.683071+ 0.0818162*Alder – 0.000868*Alder*Alder} =

Vi kan teikne ut desse samanhengane i same diagram slik det er gjort nedanfor ved å nytte ”Compute” på likninga

E.inntekt = exp{3.8477867 + 0.0666294*Alder - 0.000768*Alder*Alder +
(- 0.1647157 + 0.0151868*Alder - 0.000100*Alder*Alder)*Mann }

der ”Alder” og ”Mann” viser til variablane i datasettet (Eigentleg vil alle variablar med verdiar 15-94 kunne nyttast for Alder, og alle variablar med verdiane 0 og 1 for Mann). Resultatet er vist i diagrammet nedanfor.


Bet Eff Plott av E.inntekt etter alder for menn og kvinner
300
200
100
0
10
20
30
40
50
60
70
80
90
100
Alder

d. Modell 15 er identisk med modell 8, men er estimert utan dei 2 personane som har størst innverknad på estimatet av modell 8. Kva kan seiast om desse personane? Kva konsekvensar har det for regresjonsresultatet at personane vert utelatne?

Dei to personane som har størst innverknad på modell 8 er dei to som har høast verdi på Cook's D i modell 8. Vi har følgjande opplysningar om dei to casa som har høgaste verdi på Cook's D:

Case No	1756	1803	Kommentar
Variabelnamn			
Univ/ høyskole utd	1	0	1 = har akademisk tittel
Næring	6	2	6 = undervisning, forsking, 2 = varehandel, butikk
Ekteskapeleg status	1	2	1 = gift, 2 = aldri gift
Heiltidsarbeid	0	1	1=ja
Offentleg sektor	1	0	1=ja
Talet på arbeidstakarar i husst	2	1	#
Barn i hushaldet	0	0	#
Mors utdanning	7	7	I år
Fars utdanning	7	15	I år
E.utdanning	17	9	I år
Alder	57	19	I år
Mann	0	1	1=ja
E.innt	450	450	I 1000 kroner
HH.inntekt	450	450	I 1000 kroner
Busadtype	2	1	1 = sentrum storby, 2 = forstad storby
Kjelde til livsopphald	4	2	4 = funksjonær (ikkje leiande), 2 = arbeidar faglært
Residual E.innt Model 8	305.481	331.047	
h(i) E.innt Model 8	0.01132	0.00969	
Cook's D(i) Influence E.innt Model 8	0.02193	0.02198	

Den eine personen er ei 57 år gammal gift kvinne som arbeider innan forsking/ undervisning i offentleg sektor og har inntekt over 400.000 utan å vere i ei leiande stilling. Dette er nok ei høg inntekt dersom alt er fast lønn, men truleg har tilfeldige ekstrainntekter gjort at denne personen kom i høgaste inntektsklasse.

Den andre personen er ein 19 år gammal mann som arbeider i privat sektor i varehandel/ butikk som vanleg arbeidar og har inntekt over 400.000. Dette synest som å vere i uvanleg høg inntekt for ein så ung mann, men det er nok tenkjeleg innan visse typar salsarbeid der lønna er basert på provisjon av sal.

Desse to personane har atypiske lønnsverdiar, og får høge residualar. Men ein kan ikkje seie at verdiane er umogelege.

Kva konsekvensar har det for regresjonsresultatet at personane vert utelatne?
Vi kan studere konsekvensane ved å samanlikne regresjonsresultatet i modell 8 med resultatet i modell 15.

Variabelnamn	Modell 8 Estimate	Modell 15 Estimate	Modell 8 parameter relativt til model 15
Konstant	-24.48258	-17.43017	1.404609
Alder	3.7200401	3.5891408	1.036471
Alder*Alder	-0.035721	-0.034789	1.026790
Mann	-73.53077	-81.07991	0.906893
E.utdanning	3.2611476	2.9084876	1.121252
Heiltidsarbeid	-105.2539	-115.5403	0.910971
Offentleg sektor	17.576729	16.916805	1.039010
Alder*Mann	3.3566592	3.5562293	0.943882
Alder*Alder*Mann	-0.027637	-0.029315	0.942760
E.utdanning*Mann	1.5145129	1.8116149	0.836002
Heiltidsarbeid*Mann	23.898241	22.476676	1.063246
Alder*Heiltidsarbeid	6.6503477	7.0310364	0.945856
Alder*Alder*Heiltidsarbeid	-0.07877	-0.082802	0.951306
E.utdanning*Heiltidsarbeid	5.0927931	5.3319909	0.955139
Offentleg sektor*Heiltidsarbeid	-41.23729	-40.44984	1.019467

Ser vi på den relative storleiken av parametrane i modell 8 samanlikna med modell 15 ser vi at i modell 8 er konstanten over 40% større, effekten av Mann er nesten 10% mindre, og effekten av E.utdanning over 12% større. Effekten av interaksjonsleddet E.utdanning*Mann er nesten 17% mindre.

Den substansielle forskjellen er stor både for Mann og for E.utdanning. Det synest rimeleg å sjå dette i lys av at dei to utelatte casa var ein 19 år gammal mann med låg utdanning og ei 57 år gammal kvinne med høg utdanning. Kombinasjonen av variabelverdiar gjer at dei to casa får relativt stor innverknad på regresjonsresultatet. Fordelinga av residualen og spreiinga til residualen etter verdien av predikert Y har likevel ikkje endra seg påviseleg.

For å få betre resultat kan ein ta i bruk robust regresjon der innverknaden av uvanlege case kan vektast ned. Eit alternativ kan likevel vere å forbetre modellspesifikasjonen. Modellen kunne teke omsyn både til bustad og kjelde til livsopphald. Det kan også tenkjast at ein transformasjon av den avhengige variabelen vil gjere verknaden av utliggarar mindre.

OPPGÅVE 3 (Logistisk regresjon, vekt 0,45)
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"?

I tabellvedlegget for oppgåve 3 modell 1 finn vi at

	Estimate	Std Error	ChiSquare	Prob>ChiSq	Odds Ratio	VIF
Kvinne	1.27479645	0.1123712	128.70	<.0001	3.57797305	1.0077861
E.utdanning	0.02550665	0.0171917	2.20	0.1379	1.29054742	1.006113

I modell 1 er effekten av E.utdanning estimert til å vere 0.02550665 med ein standardfeil på 0.0171917. 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 identiske. I store utval (dvs.: fridomsgradene n-K > 120) kan vi med andre ord finne konfidensintervall for ein parameter i ein logistisk regresjonsmodell på same måten som i OLS regresjon.

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

$$b_{E.utdanning} - SE\mathbf{b}_{E.utdanning} * t_{5\%} < \beta_{E.utdanning} < b_{E.utdanning} + SE\mathbf{b}_{E.utdanning} * t_{5\%}$$

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 følgje tabell A4.1 hos Hamilton (1992:350) vil vi med meir enn 120 fridomsgrader ha at $t_{5\%}$ =1,96 (tosidig test; 1,96 = $t_{2,5\%}$ i einsidig test). Set vi inn i formelen finn vi no at

$$0.02550665 - 0.0171917 * 1.96 < \beta_{E.utdanning} < 0.02550665 + 0.0171917 * 1.96$$
$$0.02550665 - 0.03369573 < \beta_{E.utdanning} < 0.02550665 + 0.03369573$$
$$-0.0081891 < \beta_{E.utdanning} < 0.05920238$$

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.

Korleis kan ein tolke parameterestimatet for "E.utdanning"?

Parameterestimatet for E.utdanning kan tolkast på fleire måtar. Som nemnt ovanfor kan det sjåast som eit lineært tillegg i logiten for kvart år ekstra utdanning personen har. Det kan også fortelje oss kva oddsraten for å velje å vitje husflidsforretning er mellom ulike utdanningsgrupper. Sidan vi her finn at parameteren ikkje er signifikant ulik 0, kan det forsvarast å seie at effekten er lik 0. Tillegget i logiten er lik 0 og oddsraten mellom to påfølgjande utdanningskategoriar vert lik 1 (exp{0}). Det same vert den sjølvsagt når vi samanliknar dei som har mest utdanning (17år) og dei som har minst (7år).

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 $\exp[b_{E.utdanning}] = \exp[0.02550665] = 1.02583473$, 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.

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
$\Pr[Y_i = 1 \mid X_{1i}, X_{2i}, X_{3i}] = 1/ (1+ \text{Exp}[- \{-2.7133286 + 1.27479645*\text{Kvinne}_i + 0.02550665*\text{E.utdanning}_i + 0.06671985*\text{Barn i hushaldet}_i \}])$
Samanhengen mellom utdanning og sannsyn for vitje husflidsforretning studerer vi best ved hjelp av betinga effekt plott.

Korleis tolkar vi den oppgitte oddsraten for "Kvinne"?
Oddsraten for Kvinne er lik 3.57797305. Det tyder at oddsen for å vitje husflidsforretning er meir enn 3 og ein halv gong større for kvinner enn for menn.

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.

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 2 er følgjande variablar definert:

Variabel	Variabelnamn	
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.

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

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

der ε_i er feilleddet, L_i^* er estimert forventa verdi av logiten, L_i , i = 1,2,3, … ,2948 og logiten er definert ved

$$E[L_i]=\beta_0 +\beta_1 X_{1i} +\beta_2 X_{2i} +\beta_3 X_{3i} +\beta_4 X_{4i} \ .$$

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

Dette siste punktet er ikkje tatt med som ein føresetnad 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).

Finn ut om "Ekteskapeleg status" gir eit signifikant bidrag til modellen.
Det er estimert sju modellar av «Besøke hushaldsforretning». Ekteskapeleg status er dummykoda og er inkludert i 3 av dei 7 modellane med 3 variablar. Statusen "Gift" er referansekategori.

Vi kan teste om " Ekteskapeleg status" yter signifikant til modellspesifikasjonen på 3 ulike måtar. Vi kan samanlikne modellane 1og 4, eller 2 og 5, eller 3 og 6. Skilnaden mellom dei tre para av modellar er om alder ikkje er inkludert, eller om alder er inkludert som lineær eller som kurvelineær variabel.

Ved å samanlikne to modellar, ein stor modell med H fleire variablar enn i ein mindre modell med K-H parametrar, i ein sannsynsratetest (Liklihoodrate test) kan vi avgjere om dei H ekstra variablane i den store modellen samla sett yter signifikant til å forklare variasjonen i den avhengige variabelen. 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 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").

Testen er basert på nullhypotesa at regresjonskoeffesientane for dei inkluderte dummyvariablane for "Ekteskapeleg status" ikkje faktisk er 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. Til større verdi vi finn for χ^2_H til mindre sannsyn er det for at nullhypotesa kan vere rett.

Dersom Alder faktisk er kurvelineært relatert til logiten vil den sterkaste testen av "Ekteskapeleg status" vere samanlikninga av modell 3 og 6. Vi skal her gjennomføre alle tre testane. Vi har da følgjande LogLikelihoodar

Liten Modell	LogLikelihood	Stor Modell	LogLikelihood
1	-1257.923948	4	-1227.913494
2	-1220.65891	5	-1211.427021
3	-1199.138836	6	-1197.048429

Dette gir følgjande χ^2_H verdiar i testane
Modell 4 mot 1:
χ^2_H = -2* ([-1257.923948]-[-1227.913494]) = -2*(-30.010454)= 60.020908
Modell 5 mot 2:
χ^2_H = -2* ([-1220.65891]-[-1211.427021]) = -2*(-9.231889)= 18.463778
Modell 6 mot 3:
χ^2_H = -2* ([-1199.138836]-[-1197.048429]) = -2*(-2.090407)= 4.180814

Med 3 fridomsgrader vil eit kjikvadrat på 7,815 eller større gi eit signifikansnivå (= sjansen for å forkaste ei rett nullhypotese) på 0.05 eller lågare (Hamilton 1992, tabell A4.3 side 354). I testane der alder ikkje er med eller berre er inkludert som lineært element vil vi forkaste nullhypotesa om at "Ekteskapeleg status" ikkje gir signifikant bidrag til modellen. Men i det siste høvet der alder er inkludert som kurvelineært element i modellen av logiten kan vi ikkje forkaste nullhypotesa. Vi kan med andre ord ikkje avgjere om Ekteskapeleg status yter signifikant til modellen utan å ta stilling til om Alder er ein relevant variabel i 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.
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:

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**

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.**

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.

Det er ikkje spurt etter utrekning av oddsen, men vi kan ta det i alle fall.
Ut frå definisjonen:
O_i (persontypen i pkt b) = 0.17958889 / (1 - 0.17958889) = 0.21890109
Alternativt
O_i (persontypen i pkt b) = $\exp\{L(i)\}$ = $\exp\{-1.5191353\}$ = 0.21890109

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.
Oddsraten
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) =
$O_i / O_j = \exp[L(i)] / \exp[L(j)] = \exp[L(i)-L(j)]$

I dette høvet er X= E.utdanning. Det kan då argumenterast med at modellen er estimert med 17 års utdanning som høgaste verdi der alle med 18 og 19 års utdanning er koda 17. Ved å nytte same 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 2 finne at forventa verdi av logiten for person nr i er

L(i) = -4.7374066 + 1.33917673*Kvinne(i) + 0.0754151*E.utdanning(i) + 0.4663504*Barn i hushaldet(i) + 0.02919449*Alder(i)

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 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.
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.

Vi ser ut frå estimatet av modell 2 at $b_{E.utdanning} = 0.0754151$.

Med andre ord vil $\exp[b_{E.utdanning}]$ gi oss oddsraten mellom to påfølgjande verdiar av variabelen E.utdanning. Dette gjeld generelt for alle variablar.

Skriv opp formelen for å finne betinga effektplott for samanhengen mellom sannsyn og alder i modell 3.
Forventa verdi av logiten er i modell 3 estimert til
L(i) = -6.5102512 + 1.34401309*Kvinne + 0.05900844*E.utdanning + 0.25424529*Barn i hushaldet + 0.13641411*Alder - 0.0011803*Alder*Alder

Sannsynet finn vi som
Pr(Y=1) = 1/(1 + Exp{-L(i)} = 1/(1 + Exp{-(-6.5102512 + 1.34401309*Kvinne + 0.05900844*E.utdanning + 0.25424529*Barn i hushaldet + 0.13641411*Alder - 0.0011803*Alder*Alder)}

som vi også kan skrive

Pr(Y=1) = 1/(1 + Exp{6.5102512 - 1.34401309*Kvinne - 0.05900844*E.utdanning - 0.25424529*Barn i hushaldet - 0.13641411*Alder + 0.0011803*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.

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

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.

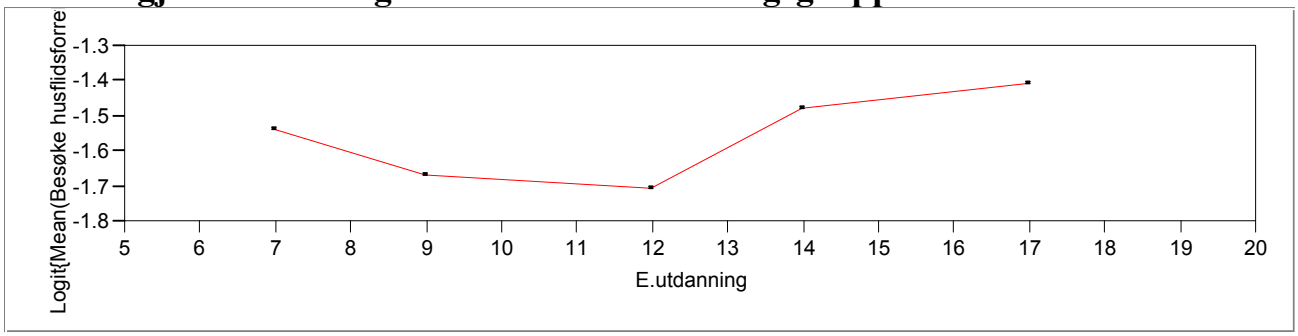
Modellen 2 har ikkje irrelevante variablar. Alle inkluderte variablar er signifikante med ein p-verdi mindre enn 0.0001.

Spørsmålet om alle relevante variablar er med er ikkje så lett å svare på. Det er først og fremst eit spørsmål om teori, og om formålet med modellen. Teori skal vi la ligge. Men dersom formålet med modellen er prediksjon må ein seie det er langt igjen til vi kan tru alle relevante variablar er inkludert.

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.

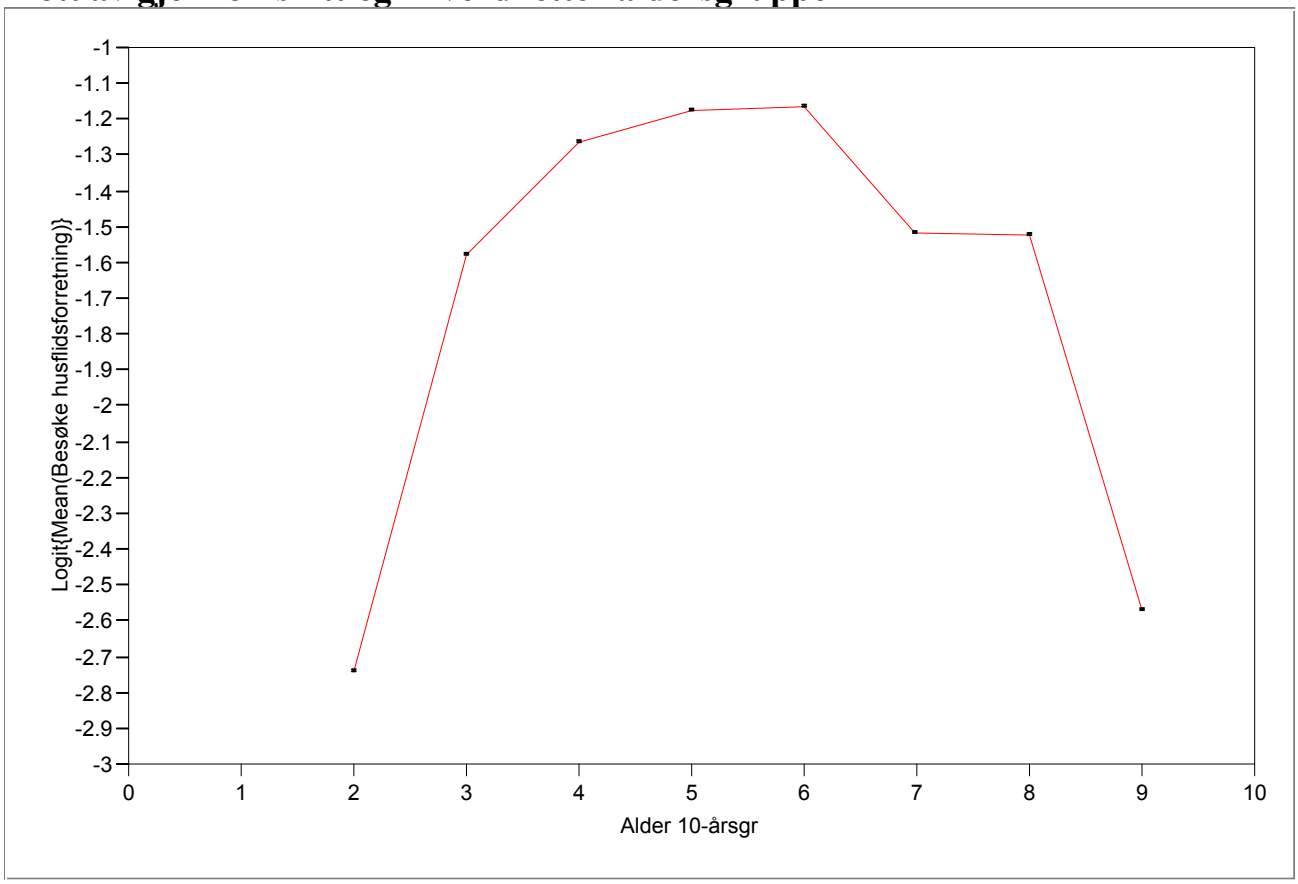
Plott av gjennomsnittleg Y-verdi etter utdanningsgrupper



Slike figurar er svært sensitive for korleis skalaen på y-aksen er framstilt. Vi må derfor sjå på variasjonsbreidda for den observerte logiten. I utgangspunktet kan logiten får verdiar frå minus uendeleg til pluss uendeleg avhengig av kor stor del av observasjonane som har Y=1 i gruppa.

Når den for E.utdanning faktisk varierer mellom –1,7 og –1,4 skjønar vi at det er heller liten variasjon mellom utdanningsgruppene i proporsjonen med Y=1. Vi ser da også i modell 7 at når E.utdanning blir inkludert som kurvelineær ved hjelp av andregradspolynom vert begge ledda (både E.utdanning og E.utdanning*E.utdanning) klart ikkje signifikante.

Plott av gjennomsnittleg Y-verdi etter aldersgrupper



Variasjonen i proporsjonen med Y=1 mellom ulike aldersgrupper er større. Logiten varierer mellom aldersgruppene frå omlag –1,2 til omlag –2,7. Heller ikkje dette er stor variasjon. Men den er tydeleg nok til at ein må konkludere med at Alder er kurvelineær i logiten. Vi finn da også både i modell 3 og 6 at begge ledda i andregradspolynomet med Alder er klart signifikante.

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	-1257.923948	4	– u/ alder	-1227.913494
2	– m/alder	-1220.658910	5	– m/alder	-1211.427021
3	– m/alder*alder	-1199.138836	6	– m/alder*alder	-1197.048429
7	– m/E.utd*E.utd	1199.114800			

E.utdanning som andregradspolynom har vi kommentert ovanfor. Sidan vi ikkje har modellar utan utdanning kan vi ikkje teste dei to ledda i polynomet samla. Men aleine er E.utdanning klart signifikant.

Alderspolynomet kan testast anten ved å samanlikne modell 6 med 4 eller ved samanlikning av modell 3 med 1. Den sterkaste testen får vi ved å samanlikne 6 med 4.

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

Modellestimering krev vidare at det ikkje er perfekt multikollinearitet eller 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.

Ein god indikator på mogelege multikollinearitetsproblem er høg verdi av variansinflasjonsfaktoren, VIF. Den er i modellane 1-7 over 2 berre der vi introduserer andregradsledd. Den er klart størst for E.utdanning og E.utdanning*E.utdanning. Dette finn vi i modell 7 og VIF har her verdiane 58,4 og 56,9. Det er derfor eit minus at vi ikkje får testa kurevelineariteten i utdanning ved sannsynsratetesten (Likelihood ratio test).

For polynomet av Alder finn vi dei største VIF verdiane i modell 6 med 36,7 og 34,5. For andregradspolynom er ikkje dette urovekkjande på nokon måte. I modell 7 er VIF verdiane for Alder og Alder*Alder mindre.

Ein kan ut frå dette konkludere med at multikollinearitet truleg ikkje er noko problem i modellane 1-6.

Diskriminering går på evna til å predikere Y-verdiane. Dersom visse x-verdiar fører til at ein predikerer Y perfekt vil vi for dummyvariablar ikkje kunne estimere nokon koeffisient og for andre variablar vil det føre til store standardfeil og upresise parameterestimat.

I krysstabellar av avhengig variabel mot dei uavhengige finn vi ei nullcelle for den dummykoda variabelen ”Uoppgitt e.status”. ”Uoppgitt e.status” predikerer Y=1 perfekt. Dette fører til at effekten av dummyen ”Uoppgitt e.status” i modellane 4-6 vert markert som ”Unstable”. For dei andre variablane er det ikkje problem verken med 0-celler eller små tal i noka celle.