EKSAMENSOPPGÅVER SVSOS3003
Vår 2004
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 med 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@svt.ntnu.no>)

Oppgåve 1 (OLS regresjon, vekt 0,5)

I ein studie av korleis dei som har opplevd kriminalitet ser på det legale systemet, har innverknaden av å ha offer for kriminalitet i familien på tilliten til det legale systemet vore studert i ei multivariat tilnærming med kontroll for verknaden av andre variable ved hjelp av OLS regresjon.
Den avhengige variabelen er "Tillit til det legale systemet". Variabelen rapporterer meininga til respondenten på ein skala frå 0 = "absolutt ikkje nokon tillit til det legale systemet" til 10 = "fullstendig tillit til det legale systemet". Fem kontrollvariablar vert introdusert i sekvens. Nokre av resultata frå analysen er inkludert i tabellappendikset til eksamensoppgåve 1.

a)
Drøft relasjonen mellom "å ha offer for kriminalitet i familien" og "tilliten til det legale systemet" slik det kjem til uttrykk i desse regresjonsanalysane

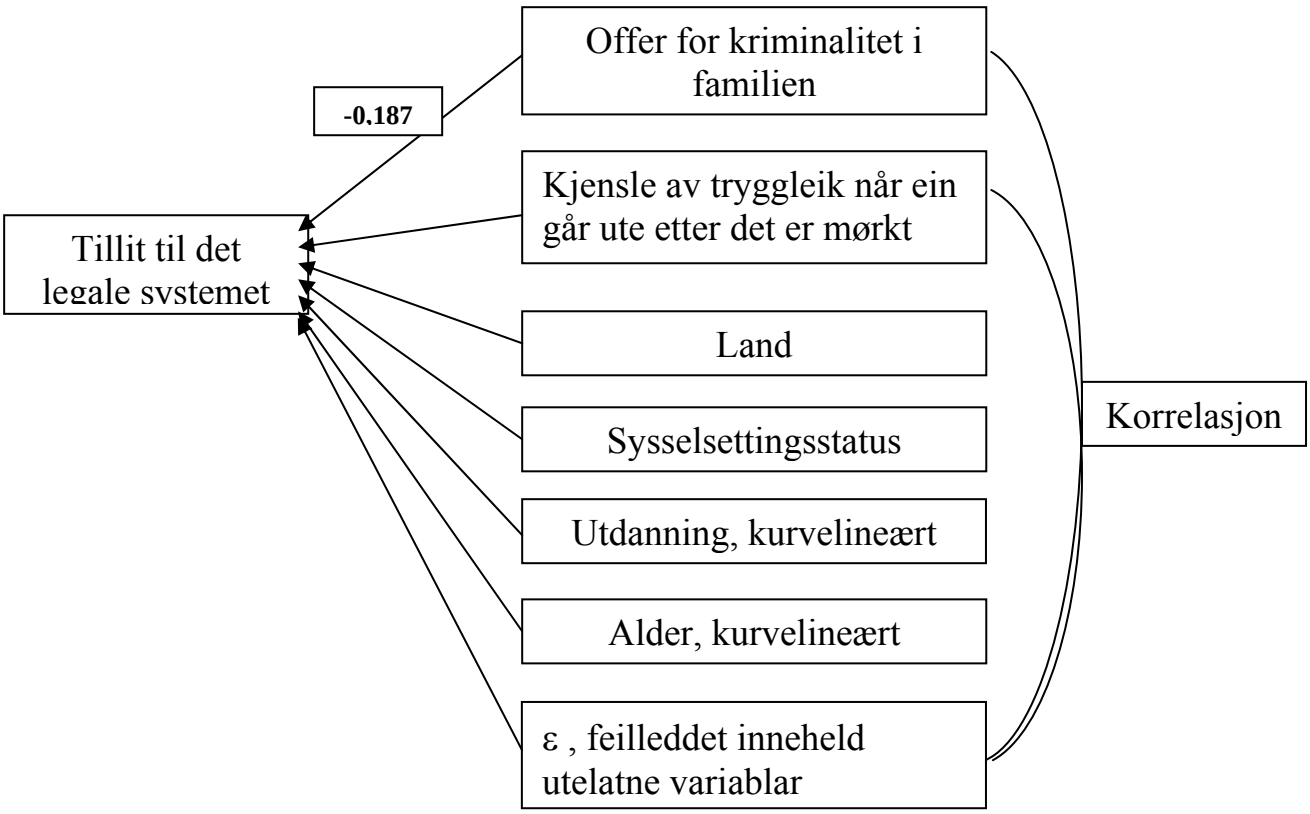
I den siste og mest omfattande modellen ser vi at på 10-punktsskalaen over tillit til det legale systemet minkar tilliten med 0,187 poeng dersom personen har eit offer for kriminalitet i familien etter kontroll for effekten av kjensle av kunne gå trygt etter det er mørkt, land ein bur i, sysselsetjingsstatus, utdanning og alder. Effekten er signifikant ulik null med sannsyn mindre enn 0,01. Opplysningar om samanhengen i dei ulike modellane kan presenterast som nedanfor:

Dependent Variable: Trust in the legal system	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6
(Constant)	5,568	6,014	5,425	5,493	5,020	5,579
Victim of crime in IP's family	**-,187**	**-,099**	**-,142**	**-,157**	**-,199**	**-,187**
Safe after dark		-,431	-,144	-,131	-,136	-,146
Unsafe after dark		-,978	-,504	-,474	-,445	-,457
Very unsafe after dark		-1,803	-1,225	-1,166	-1,088	-1,121
Spain			-,812	-,789	-,714	-,685
Sweden			,858	,852	,913	,908
Norway			1,071	1,055	1,049	1,044
selfempl				-,025	-,007	,005
notempl				-,187	-,049	-,191
Education in years					-,003	,013
Education in years squared					,003	,002
Age in years						-,034
Age in years squared						,000089

I modell 1 ser vi at den bivariate samanhengen mellom dei to variablane er på -0,187, nett det sam som i modell 6 etter kontroll for 5 andre variablar. I modell 2 kjem den opp i -0,099 ved introduksjon av variabelen "kjensle av tryggleik når ein er ute og går etter det er mørkt" og den er nede i -0.199 i modell 5. I modell 2 er den likevel ikkje signifikant ulik null med nivå 0,05 på testen.

I modell 1 er alle andre variablar enn "offer for kriminalitet i familien" ekskludert frå modellen. Dersom ekskluderte variablar korrelerer med inkluderte variablar samtidig som dei har verknad på den avhengige variabelen er dei relevant for modellen og skal inkluderast. Den relativt store endringa i regresjonskoeffisienten (frå -0,187 til -0,099) ved kontroll for kjensla av tryggleik viser at det anten er ein viss korrelasjon mellom dei to variablane ("kjensle av tryggleik …" og "offer for kriminalitet i familien") eller at dei kvar for seg og på ulikt vis korrelerer med ekskluderte variablar.

Toleransen til "offer for kriminalitet i familien" er heile tida høg og sjølv i modell 6 er den over 0,96. Dette viser at det ikkje er korrelasjonar mellom "offer for kriminalitet i familien" og nokon av dei variablane som er brukt i modellane her som er opphavet til endringane i effekten. Den rimelegaste forklaringa er at det finst viktige ekskluderte variablar som ikkje er inkludert i nokon av dei 6 modellane.



b)

Finn konfidensintervallet for regresjonskoeffisienten til "å ha offer for kriminalitet i familien" med signifikansnivå 0,01. Test om "sysselsettingsstatus" gjev ei signifikant yting til modellen

Konfidensintervall

I modell 6 er $b_{Victim\ of\ crime}$, effekten av "å ha offer for kriminalitet i familien", oppgitt til å vere **-0,187** med ein standardfeil på 0,062. Dersom vi kan gå ut frå at feilledda er normalfordelte vil eit 99% konfidensintervall (1% signifikansnivå) vere gitt ved

$$b_{Victim\ of\ crime} - SE_{Victim\ of\ crime} * t_{1\%} < \beta_{Victim\ of\ crime} < b_{Victim\ of\ crime} + SE_{Victim\ of\ crime} * t_{1\%}$$

der b er den estimerte regresjonskoeffisienten, SE er standardfeilen til regresjonskoeffisienten og $t_{1\%}$ er den kritiske verdien i t-fordelinga i ein tosidig test med signifikansnivå 0,01. I følgje tabell A4.1 hos Hamilton (1992:350) vil vi med meir enn 120 fridomsgrader ha at $t_{1\%} = 2{,}576$ (tosidig test). Set vi inn i formelen finn vi no at

$$-0{,}187 - 0{,}062 * 2{,}576 < \beta_{Victim\ of\ crime} < -0{,}187 + 0{,}062 * 2{,}576$$
$$-0{,}187 - 0{,}159712 < \beta_{Victim\ of\ crime} < -0{,}187 + 0{,}159712$$

dvs.

$$-0{,}347 < \beta_{Victim\ of\ crime} < -0{,}0273$$

I 99 av hundre utval vil vi finne eit estimat av $\beta_{Victim\ of\ crime}$ som ligg mellom -0,347 og -0,0273.

Sysselsettingsstatus

Det skal testast om "sysselsettingsstatus" gjev ei signifikant yting til modellen. Det er det ikkje eksplisitt spesifisert noko nivå for denne testen. Men det synest rimeleg å nytte same nivå som i første delen av spørsmålet.

Sysselsettingsstatus er første gong introdusert i estimeringa i modell 4. Ein test av om sysselsettingsstatus gir ei signifikant yting til modellen er det same som å teste om det er signifikant skilnad mellom modell 3 og modell 4.

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

$$F^{H}_{n-K} = \frac{\frac{RSS_{[K-H]} - RSS_{[K]}}{H}}{\frac{RSS_{[K]}}{n-K}}$$

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 : "Ingen effekt av nye variablar" er rett). I formelen er $RSS_{[K]}$ kvadratsummen til residualane i den mest omfattande modellen med K parametrar (eller K-1 variablar) og $RSS_{[K-H]}$ er kvadratsummen til residualane i modellen som manglar dei H variablane som skal testast. Vi forkastar null-hypotesa om at alle koeffisientane til dei H ekstra variablane er null med signifikansnivået α dersom F^{H}_{n-K} er større enn den kritiske verdien for signifikansnivået α i F-fordelinga med H og (n-K) fridomsgrader.

I tabellappendikset finn vi at

Model		Sum of Squares	df
1	Regression	49,511	1
	Residual	44253,295	7389
	Total	44302,806	7390
2	Regression	1708,098	4
	Residual	42594,709	7386
	Total	44302,806	7390
3	Regression	5198,063	7
	Residual	39104,744	7383
	Total	44302,806	7390
4	Regression	5254,469	9
	Residual	39048,337	7381
	Total	44302,806	7390
5	Regression	5680,968	11
	Residual	38621,838	7379
	Total	44302,806	7390
6	Regression	5782,884	13
	Residual	38519,922	7377
	Total	44302,806	7390

I lesing av slike tabellar kan det vere greitt å hugse at talet på fridomsgrader (df) for TSS (total sum of squares) er lik n-1. Vi ser da at n for dei modellane vi estimerer er 7391 (n-1=7390).

Samanliknar vi modell 3 og 4 ser vi at
H = 2
K = 10
n-K= 7391 - 10 = 7381
$RSS_{[K-H]}$ = 39104,744
$RSS_{[K]}$ = 39048,337
Vi finn at F^{2}_{7381} = 5,3282. Sidan 1% kritisk verdi i F-fordelinga med 2 og 7381 fridomsgrader er 4,61 (Hamilton 1992, tabell A4.2, s 353) vil vi forkaste nullhypotesa om at variabelen sysselsettingsstatus ikkje yter til å forklare variasjonen i tillit til det legale systemet. Resonnementet er slik: Når nullhypotesa er korrekt er det eit sannsyn som er 0,01 eller mindre for å finne ein F (med 2 og uendeleg mange fridomsgrader)

som er 4,61 eller større. Sidan vi fann ein F på 5,32 konkluderer vi med at den er for stor, eller sannsynet for å finne ein så stor F er for lågt, til at vi kan tru på nullhypotesa.

Tilsvarande finn vi i ei samanlikning av modell 4 og 5 at
H=2
K=12
n-K= 7391 - 12 = 7379
$RSS_{[K-H]}$ = 39048,337
$RSS_{[K]}$ = 38621,838
Dette gir F^2_{7379} = 40,747, slik at utdanning gir eit signifikant bidrag til modellen ved kurvelineær inklusjon.

c)
Lag eit betinga effektplott av effekten til land (country) i modell 6.

Meininga med å lage eit slikt plott er utan tvil å samanlikne dei fire landa i granskinga. Det skulle då vere nok å sjå på den relative storleiken av effektane som dei ulike landa har. I modell 6 finn vi at

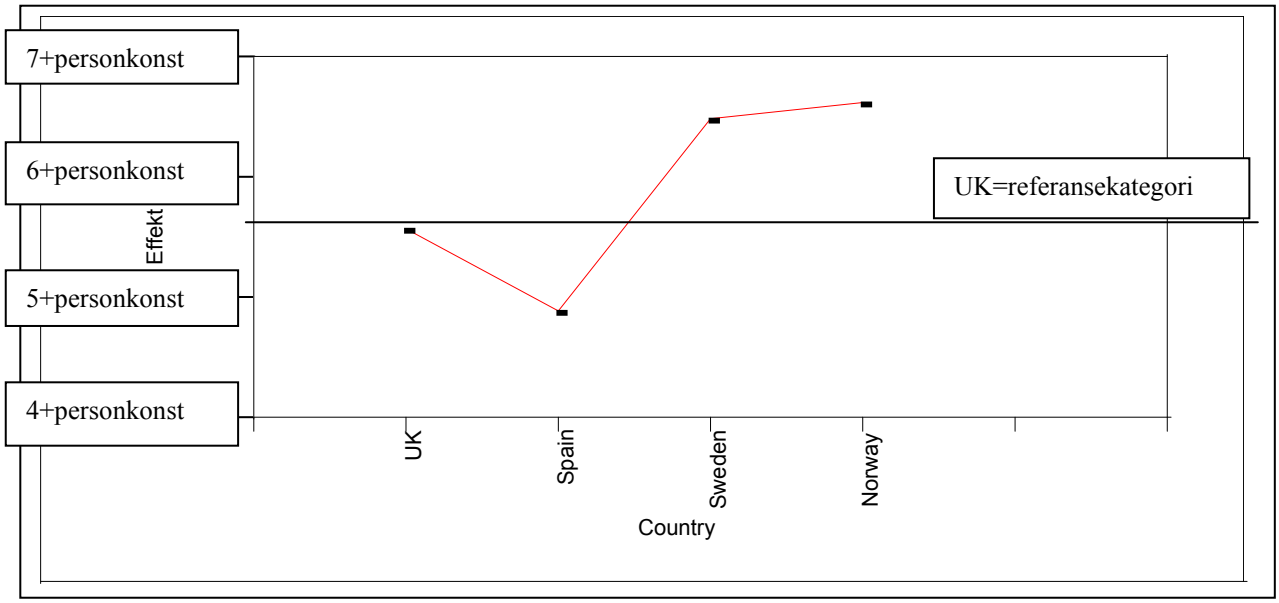
Predikert Y = 5,579 -0,187*Victim of crime in IP's family -0,146*Safe after dark -0,457*Unsafe after dark -1,121*Very unsafe after dark -0,685*Spain +0,908*Sweden +1,044*Norway +0,005*selfempl -0,191*notempl +0,013*Education in years +0,002*Education in years squared -0,034*Age in years +0,000089*Age in years squared

I denne likninga finn vi to typar ledd: dei som er påverka av kva land personane er busett i og dei som ikkje er det. Det er verd å hugse at "konstanten" er påverka av alle dummykoda variablar i modellen inklusiv land Dei ledda som ikkje er påverka av land kan vi setje lik konstanten "personkonst". Vi får da ein enklare samanheng:

Predkert Y = **5,579 -0,685*Spain +0,908*Sweden +1,044*Norway** + personkonst, der
personkonst = -0,187*Victim of crime in IP's family -0,146*Safe after dark -0,457*Unsafe after dark -1,121*Very unsafe after dark +0,005*selfempl -0,191*notempl +0,013*Education in years +0,002*Education in years squared -0,034*Age in years +0,000089*Age in years squared

Effekten av land på tillit til det legale systemet vil dermed vere som følgjer:

Effekten av å bu i UK for ein gitt person:	UK effekt = personkonst +5,579
Effekten av å bu i Spania for ein gitt person:	ES effekt = personkonst +4,894
Effekten av å bu i Sverige for ein gitt person:	SE effekt = personkonst +6,487
Effekten av å bu i Norge for ein gitt person:	NO effekt = personkonst +6,623



d)
Formuler den fullstendige modellen som er estimert

Når ein modell skal formulerast trengst det gjerast greie for tre typar element:

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

Med utgangspunkt i data frå Storbritannia, Spania, Sverige og Norge samla inn i 2002 gjennom "European Social Survey" (ESS) er følgjande "tekniske" variablar definert for bruk i modellbygging:

Y	Trust in the legal system
X_1	**Victim of crime in IP's family**
	Feeling of safety walking alone in local area after dark
X_2	Safe after dark
X_3	Unsafe after dark
X_4	Very unsafe after dark
	Country

X_5	Spain
X_6	Sweden
X_7	Norway
	Employment status
X_8	selfempl
X_9	notempl
	Education
X_{10}	Education in years
X_{11}	Education in years squared
	Age
X_{12}	Age in years
X_{13}	Age in years squared

I alle landa som er med i ESS er det gjort tilfeldige utval frå befolkninga slik at det kan trekkjast konklusjonar om befolkninga i kvart land for seg.

Oppgåva er å forklare variasjonen i tilliten til det legale systemet i dei fire landa. Vi føreset at det er eit lineært eller kurvelineært samband mellom Y og dei definerte X-variablane slik at vi kan skrive

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

når vi lar i gå over heile populasjonen. Lar vi k=0, 1, 2, ... ,13, 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", feilleddet, er ein variabel som fangar opp dei faktorane som ikkje har vore observert saman med reint tilfeldig støy i målinga av Y_i.

I dei fire landa er det i alt 7816 personar som har vore intervjua. Dei fordeler seg slik på dei fire landa:

Spain	1729
United Kingdom	2052
Norway	2036
Sweden	1999
Total	**7816**

Det manglar opplysningar på mange personar på ulike variablar. Etter listevis utelating av manglande data på ein eller fleire av variablane som er definert ovanfor er det igjen 7391 case som kan nyttast i analysen. Det er ingen haldepunkt for å tru anna enn at fråfallet er reint tilfeldig i høve til svar som er registrert på den avhengige variabelen. Listevis utelating er da ein fullgod prosedyre så lenge ein sit igjen med nok case til å gjennomføre analysane. Det gjer vi i dette høvet.

Tillit til det legale systemet er ei generell og truleg rimeleg stabil haldning til ein grunnleggjande samfunnsinstitusjon. Haldningar kan ein tenkje seg vil bli forma

gjennom oppvekst og erfaringar i den sosiale posisjonen kvart individ har. Den kausale strukturen er dermed at verknader går frå konkrete erfaringar (å ha eit offer for kriminalitet i familien) til tilliten til systemet. Indikatorar for lokalisering i den sosiale strukturen (alder, utdanning, sysselsettingsstatus, land) vil kunne nyttast som generelle indikatorar på kva erfaringar som har akkumulert seg. Andre haldningsvariablar (som t.d. kjensle av tryggleik når ein er ute og går etter at det er mørkt) kan ikkje ha same kausale statusen. Dei kan likevel takast inn i ein modell som kontrollvariablar. Her kan ein tenkje seg at kjensla av tryggleik kan stå for type personlegdom i form av indikasjon på generell tillit til livet. Effekten av dei andre variablane kan da estimerast netto etter at effekten av grad av generell tillit til livet er kontrollert for.

Modellen kan skrivast $y_i = E[y_i] + \varepsilon_i$

Dette tyder at $E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} +...+ \beta_{13} x_{i,13}$

($E[y_i]$ les vi som forventa verdi av y_i)

Vi finn eit OLS estimat av denne modellen som dei b-verdiane i

$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} +...+ b_{13} x_{i,13}$

($\hat{y}_i$ les vi som estimert eller "predikert" verdi av y_i eller berre y-hatt)

som minimerer kvadratsummen av residualane

$$RSS = \Sigma_i(y_i - \hat{y}_i)^2 = \Sigma_i e_i^2$$

Desse estimata er forventningsrette og variansminimale med kjent samplingfordeling dersom vi kan dokumentere at følgjande føresetnader er rette:

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 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 same varians for alle case,
  dvs: $\varepsilon_i \sim N(0, \sigma^2)$ for alle i .

e)

Drøft i kva grad føresetnadene for ein OLS-regresjon er stetta

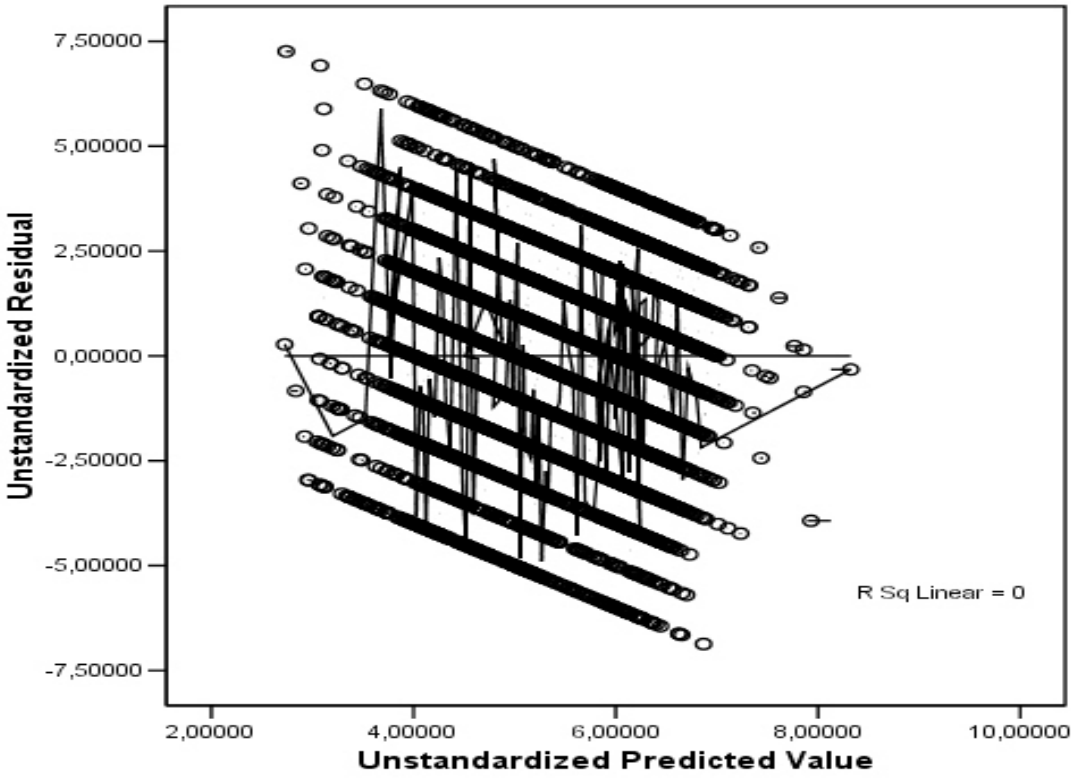
Ikkje alle føresetnadene kan testast. Vi kan ikkje seie noko om alle relevante variablar er med. Vi kan ikkje seie noko om det kan vere målefeil i x-variablane, og heller ikkje kan vi avgjere om forventa verdi til feilleddet faktisk er 0.

I spesifikasjonskravet kan vi vurdere funksjonsforma (linearitet i parametrane) og om irrelevante variablar er inkludert. To variablar, utdanning og alder, er inkludert som andregradskurver. Dei andre variablane er dikotome dummykoda variable og kan ikkje inkluderast på anna måte enn som lineære ledd. Modellane er med andre ord lineære i parametrane reint formelt. Men om restleddet ikkje er normalfordelt kan dette mellom anna skuldast ikkje linearitet. Dette ser vi på nedanfor.
I modell 6 er alle variablar signifikante utanom førstegradsleddet i utdanningsvariabelen og dummyen for sjølvsysselsett i sysselsettingsstatusvariabelen, men som vist ovanfor er begge desse variablane signifikante på 1% nivå. Vi har med andre ord ikkje irrelevante variablar i modellen.

Av dei to Gauss-Markov krava som kan testast er ikkje kravet om fråvær av autokorrelasjon relevant sidan vi her har eit tilfeldig utval på landnivå. Kravet om homoskedastisitet kan vi studere gjennom plottet av residualane mot predikert verdi:

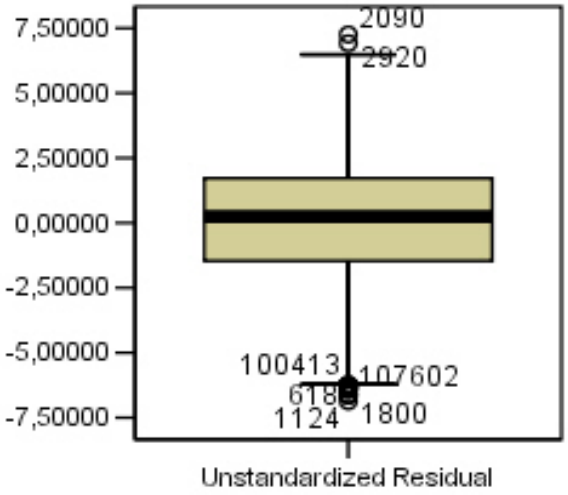
Avhengig variabel kan berre ha 11 ulike verdiar. Den innebygde heteroskedastisiteten som følgjer av at det er avkorta variasjon i den avhengige variabelen er tydeleg. Residualane kjem ut i klare linjer. Systematisk variasjon i **spreiinga av residualen** etter storleiken på ein eller fleire x-variablar er indikasjon på heteroskedastisitet. Den innlagte taggete linja[1] gir oss ei "kurve" eller "linje" som gjennom lokal vekting tilpassar seg observasjonane maksimalt. I dette høvet kan den kanskje indikere avtakande spreiing hos residualane med aukande x-verdiar. Det er likevel ikkje tydeleg om dette er noko meir enn det som følgjer av den innebygde heteroskedastisiteten.

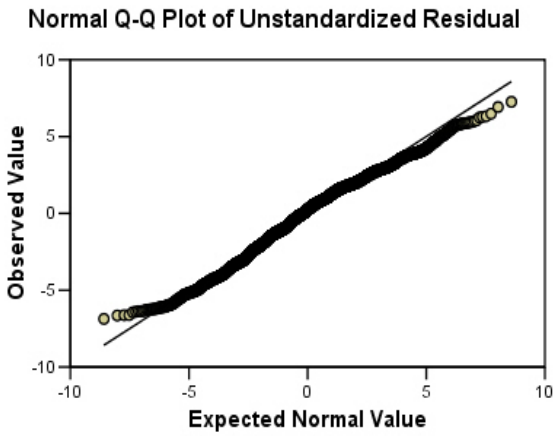
[1] Linja vert laga i "Chart Editor" ved å velge "Add Chart Element" og "Interpolation Line" og til slutt "Spline"



Heteroskedastisitet fører til at standardfeilen til regresjonskoeffisienten vert skeivt estimert og t- og F-testar vert upålitelege. I dette høvet er ikkje avviket frå det ideelle større enn at vi i alle fall eit stykke på veg vil tru på testane. Dersom testresultatet er viktig kan det likevel vere verdt å gå vidare med meir formelle testar for homoskedastisitet. Hardy (1993:60-61) gir ein kort omtale av to slike testar som kan nyttast med dummyvariablar.

Det siste kravet, normalfordelte feilledd, sjekkar vi gjennom inspeksjon av eit boks-plott av residualane og eit kvantil-normal plott for residualane.





Begge diagramma viser eit lite avvik frå normalfordelinga. Halane er lettare enn i ei tilsvarande normalfordeling, og med den avgrensinga vi har i variasjonen i y er det heller ikkje mogeleg å finne utliggjarar.

Men avviket frå normalfordelinga er ikkje stort nok til å forkaste testane som ubrukbare. Likevel, både det vesle innslaget av heteroskedastisitet og det vesle avviket frå normalfordelinga bør gjere at vi er litt varsame med konklusjonane.

f)
Drøft mogeleg indikasjonar på problem relatert til multikollinearitet og case med innverknad

Ein sikker indikator på at vi har multikollinearitet er låg toleranse. I tabellane til oppgåver 1 er toleransen oppgitt. Låg toleranse er ikkje problematisk dersom det er eit resultat av at vi har introdusert kvadratledd eller interaksjonsledd for å få ein betre spesifisert modell. Vi ser at dei to variablane ”Education in year” og ”Age in years” begge er inkludert som andregradspolynom. Toleransen er tilsvarande låg, mindre enn 0,06 for alle fire ledda. Konsekvensen av dette er at vi ikkje kan stole på t-testen for kvart einskild ledd i polynomet. Vi må nytte F-testen for å teste kvart polynom for seg som ei gruppe på 2 variablar. Dersom polynomet samla gir ei signifikant yting til modellen er det uinteressant kva for ein lekk i polynomet som regresjonsprosedyren og t-testen plukkar ut som viktig.

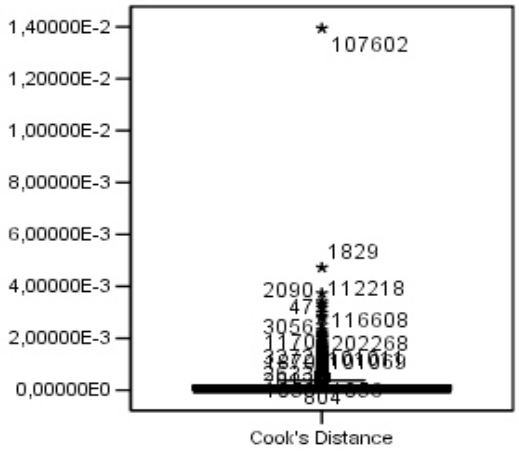
Model 6	variables	tolerance
	Victim of crime in IP's family	,965
	Safe after dark	,722
	Unsafe after dark	,718
	Very unsafe after dark	,809
	Spain	,617
	Sweden	,642
	Norway	,620
	selfempl	,928
	notempl	,610
	Education in years	,054
	Education in years squared	,059
	Age in years	,030
	Age in years squared	,027

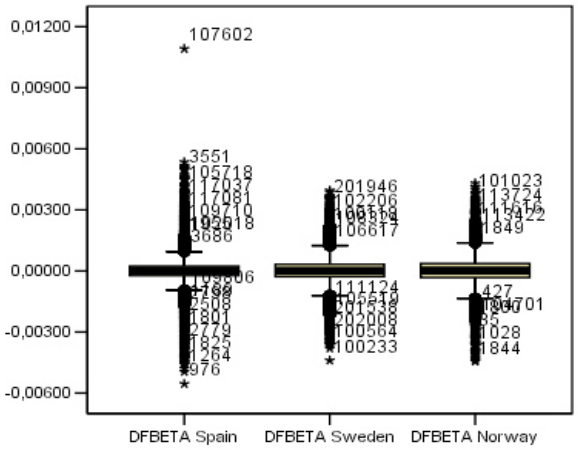
Det er ovanfor i punkt 1b vist at utdanning gir ei signifikant yting til å forklare variasjonen i Y. Vi kunne her gjere same testen for alder. Men vi ser og at begge ledda i alderspolynomet er signifikante med nivå 0,05 i følgje t-testen. Da kan ikkje F-testen vise noko dårlegare resultat. F-testen vil vere overflødig.

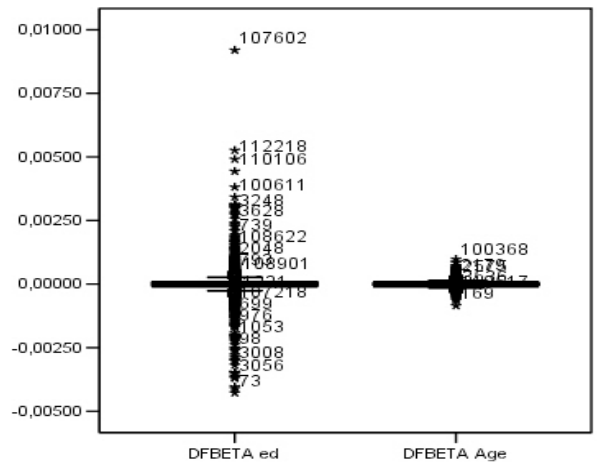
I modell 6 fører ikkje multikollinearitet til problem.

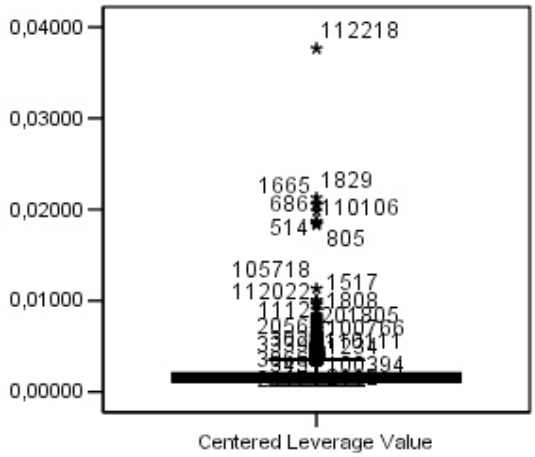
Eit case kan ha potensiale for innverknad (hatt-observatoren) og det kan ha faktisk innverknad (Cook's D, DFBETAS, store residualar).

I boksplotta over Cook's distance, DFBETAS, og den sentrerte leverage er det 2 case som skil seg ut med verdiar markert større enn dei andre. Det er casa 107602 med store verdiar for Cook's D og store DFBETAS for variablane Spain og Education in years (ed).









Ser vi på den sentrerte leverage er det eit anna case som merkar seg ut: 112218. Dette er da eit case som ikkje kjem fram i dei andre observatorane. Det har dermed ingen faktisk innverknad på resultatet. Den faktiske verdien er også for låg til at det kan ha særleg sterk innverknad i noko fall.

Vi har i punkt 1e notert at vi ikkje har utliggjarar i Y. Dermed er det svært vanskeleg å få uvanleg store residualar også. I heile materialet finn vi berre 3 case med residual som er større enn 3 standardavvik frå gjennomsnittet.

Case Number /id no	Std. Residual	Trust in the legal system	Predicted Value	Residual
854 /2090	3,029	10	3,08	6,923
1188 /2920	3,177	10	2,74	7,259
2330 /107602	-3,006	0	6,87	-6,868

I tabellen over desse casa gitt i eksamensoppgåva manglar det diverre kryssreferanse mellom case nummer og identitetsnummer. Dermed kan vi ikkje identifisere case nummer 2330 som den personen som med identitetsnummer 107602 har vist seg fram med store verdiar på Cook's D, og DFBETAS for "Spain" og "ed".

Konklusjonen bør likevel bli at det er eitt case som kan vere interessant å sjå nærmare på for å vurdere storleiken på innverknaden det har. Det er case 107602. Ein regresjon utan dette caset samanlikna med den som er rapportert her vil vise om det har substansiell verknad på regresjonskoeffisientane.

Oppgåve 2 (Logistisk regresjon, vekt 0,5)

I samband med studien av korleis dei som har offer for kriminalitet i familien ser på det legale systemet, studerte ein også innverknaden av å ha offer for kriminalitet i familien på opplevinga av å vere svært utrygg når ein var ute og gikk aleine etter det vart mørkt. Samanhengen vart studert i ei multivariat tilnærming med kontroll for verknaden av andre variable ved hjelp av logistisk regresjon.

Den avhengige variabelen er "Kjensle av å vere svært utrygg når du er ute og går aleine etter det vert mørkt". Variabelen er koda 1 for dei som svarar "Svært utrygge" på spørsmål etter kor vidt dei har "Kjensle av å vere svært utrygg når du er ute og går aleine etter det vert mørkt". Dei som svarar noko anna får koden 0. Det vert nytta listevis utelating ved manglande opplysningar. Åtte kontrollvariablar vert introdusert. Nokre av resultata frå analysen er inkludert i tabellappendikset til eksamensoppgåve 2.

a)
Drøft relasjonen mellom "å ha offer for kriminalitet i familien" og "kjensle av å vere svært utrygg når du er ute og går aleine etter det vert mørkt" slik det kjem til uttrykk i desse regresjonsanalysane.

I den logistiske regresjonen av kjensle av å vere utrygg når ein er ute og går etter det er mørkt finn vi at koeffisienten for variabelen victim (å ha eit offer for kriminalitet i familien) i likninga for logiten er 0,48 etter kontroll for verknaden av dei andre variablane. Denne koeffisienten er signifikant med ein Wald-observator på 18,275 (Wald-observatoren er kjikvadratfordelt med 1 fridomsgrad). Exp (0,48) gir oddsraten for å seie ja på spørsmålet om ein kjenner seg svært utrygg når ein er ute og går etter det er mørkt. Verdien exp(0,48) = 1,616 viser at den som har offer for kriminalitet i familien har 61,6% større sjanse for å seie ja på spørsmålet enn ein som ikkje har offer for kriminalitet i familien.

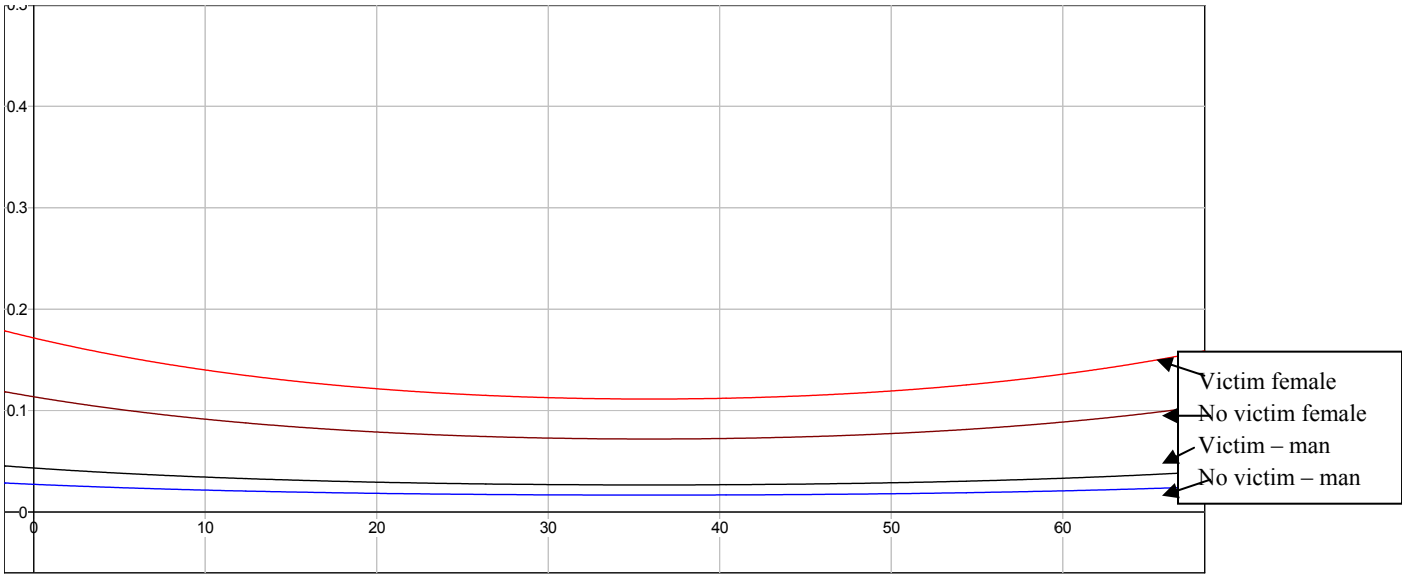
Dersom vi ønskjer å sjå på sannsynet for å kjenne seg svært utrygg er det viktig å notere seg at i logistiske modellar er alle effektar interaksjonseffektar mellom variablar inkludert i modellen. Slike samanhengar studerer ein best i grafar av sannsynet betinga av ulike verdiar på dei andre variablane. I eksamenssituasjonen tar det for mye tid å rekne på det. Ein ventar ikkje å finne slike utrekningar. Men nedanfor er det gitt eit eksempel.

Den substansielle konklusjonen ein kunne dra basert på føreliggande modellestimat kunne til dømes vere at kriminalitet ikkje berre påverkar livskvaliteten til det einskilde offeret, men i høg grad også familiemedlemmer til offeret. Dette gjeld uavhengig av kva land ein bur i, kva type lokalitet ein bur i, kor lenge ein har budd på staden, kva kjønn ein har, kor gammal ein er, om ein bur saman med ein partnar, kor mye

utdanning ein har og kva sysselsettingsstatus ein har. Det er eit resultat som synest robust i høve til alternative forklaringar som katolsk/ protestantisk kultur (Spania vs dei andre landa), by/ bygdesamfunn, einsleg, kvinne, gammal, og kunnskapsrik.

Test blocks	**2LogLike-lihood of Block**	Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Block 0	-3556,912	Constant	-,233	,534	,190	1	,663	,792
Block 1	-3545,014	victim	,480	,112	18,275	1	,000	1,616
Block 2	-3194,117	yrlvdae	-,004	,003	1,518	1	,218	,996
		eduyrs	-,165	,027	36,991	1	,000	,848
		female	1,519	,126	144,972	1	,000	4,570
		liveWithPartner	-,159	,111	2,046	1	,153	,853
Block 3	-3150,129	selfempl	,336	,256	1,713	1	,191	1,399
		notempl	,430	,131	10,804	1	,001	1,537
Block 4	-3129,757	age	-,028	,016	3,071	1	,080	,973
		age2	,00039	,000	6,388	1	,011	1,000
Block 5	-3029,321	suburb	-,525	,168	9,747	1	,002	,592
		town	-,622	,150	17,172	1	,000	,537
		village	-1,554	,183	72,302	1	,000	,211
		countryside	-1,455	,311	21,892	1	,000	,234
Block 6	-2870,363	Spain	-2,274	,371	37,582	1	,000	,103
		Sweden	-1,722	,510	11,426	1	,001	,179
		Norway	-1,874	,624	9,010	1	,003	,154
Block 7	-2849,161	eduInSpain	,131	,032	16,837	1	,000	1,140
		eduInSweden	,021	,045	,208	1	,648	1,021
		eduInNorway	,017	,052	,107	1	,743	1,017

Tolking av effekten i sannsynsskalaen krev at ein ser denne i samanheng med effekten av andre variablar. Ein må setje inn rimelege verdiar for dei andre variablane presentere det i betinga effekt plott der vi kan sjå korleis det å kjenne seg svært utrygg verkar samanlikna med det å ikkje gjere det. Nedanfor ser vi korleis dette verkar for menn og kvinner som etter 20 års utdanning arbeider og lever aleine i storbyar i Storbritannia der dei har budd i 10 år. Vi ser at effekten er større for kvinner enn for menn. Dette illustrerer korleis alle effektar er interaksjonseffektar i logistiske modellar.



y=1/(1+exp(-(-0.233+0.48×1-0.004×10-0.165×20+1.519×0-0.159×0-0.028×x+0.00039×x×x)))
y=1/(1+exp(-(-0.233+0.48×0-0.004×10-0.165×20+1.519×0-0.159×0-0.028×x+0.00039×x×x)))
y=1/(1+exp(-(-0.233+0.48×1-0.004×10-0.165×20+1.519×1-0.159×0-0.028×x+0.00039×x×x)))
y=1/(1+exp(-(-0.233+0.48×0-0.004×10-0.165×20+1.519×1-0.159×0-0.028×x+0.00039×x×x)))

b)

Finn konfidensintervallet for regresjonskoeffisienten til "å ha offer for kriminalitet i familien" med signifikansnivå 0,01. Test om "sysselsettingsstatus" gjev ei signifikant yting til modellen

Konfidensintervall

Variabelen "å ha offer for kriminalitet i familien" kan vi kalle victim.
Det er knytt uvisse til talet som gir oss effekten av victim på logiten eller sannsynet for å kjenne seg utrygg når ein går ute aleine etter det er mørkt. Kor stor uvissa er kan vi få vite noko om ved å finne eit konfidensintervall for effekten. I tabellvedlegget for oppgåve 3 modell 1 finn vi at regresjonskoeffisienten for victim er 0,480 med eit standaravvik på 0,112

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.: når fridomsgraden n-K > 120) kan vi med andre ord finne konfidensintervall for ein parameter i ein logistisk regresjonsmodell på same måten som i OLS regresjon.

For å få eit testnivå på 0,01 må konfidensintervallet dekke den rette parameterverdien med sannsyn 0,99. Eit 99% konfidensintervall for effekten av victim er da gitt ved

$$b_{victim} - SE\mathbf{b}_{victim} * t_{1\%} < \beta_{victim} < b_{victim} + SE\mathbf{b}_{victim} * t_{1\%}$$

der b er regresjonskoeffisienten, SE er standardfeilen til regresjonskoeffisienten og t er fraktilen i t-fordelinga i ein tosidig test med signifikansnivå 0,01. I følgje tabell A4.1 hos Hamilton (1992:350) vil vi med meir enn 120 fridomsgrader ha at $t_{1\%} = 2{,}576$ i ein tosidig test. Set vi inn i formelen finn vi no at

$0{,}48 - 0{,}112 * 2{,}576 < \beta_{victim} < 0{,}48 + 0{,}112 * 2{,}576$
$0{,}48 - 0{,}288512 < \beta_{victim} < 0{,}48 + 0{,}288512$
$0{,}191488 < \beta_{victim} < 0{,}768512$

I 99 av 100 granskingar av spørsmålet om kven som kjenner seg utrygg når ein går ute aleine etter det er mørkt vil konklusjonen bli at å ha eit offer for kriminalitet i familien aukar logit med mellom 0,19 og 0,77 logit-einingar.

Ein kan og seie at oddsen for å kjenne seg utrygg aukar med mellom 21,1 og 115,7% dersom ein har eit offer for kriminalitet i familien ($\exp\{0{,}191488\} = 1{,}211$ og $\exp\{0{,}768512\} = 2{,}157$).

Sysselsettingsstatus

Vi skal teste om "sysselsettingsstatus" gjev ei signifikant yting til modellen. Det er det ikkje eksplisitt spesifisert noko nivå for denne testen. Men det synest rimeleg å nytte same nivå som i første delen av spørsmålet.

Sysselsettingsstatus er inkludert ved å dummykode variabelen. Referanse kategorien er "employed". Inkluderte kategoriar er "self-employed" og "not in paid work".

Ved å samanlikne to modellar, ein stor modell med H fleire variablar enn i ein mindre modell med K-H parametrar, i ein sannsynsratetest (Likelihoodratio 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

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

der $\mathcal{L}$ står for Likelihooden, K er talet på parametrar i den største modellen og H er 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 = 2, talet av inkluderte dummyvariable for "sysselsettingsstatus").

Testen er basert på nullhypotesa at regresjonskoeffisientane for dei inkluderte dummyvariablane for "sysselsettingsstatus" eigentleg er lik 0. Dersom denne hypotesen 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.

I tabellen ovanfor finn vi at skilnaden mellom blokk 2 og 3 er variabelen sysselsettingstatus inkludert med 2 dummyvariable. Vi ser der at 2*Loglikelihooden for Block 2 er -3194,117. For Block 3 er den -3150,129. I testen for blokk 3 er det 8 parametrar (K=8), i testen for blokk 2 er det 2 færre parametrar (H=2).

Testobservatoren blir da

$$\chi^2_2 = -2\{\log_e \mathcal{L}_6 - \log_e \mathcal{L}_8\} = -(2\log_e \mathcal{L}_6) + (2\log_e \mathcal{L}_8) = -(-3194{,}117) + (-3150{,}129)$$
$$= 3194{,}117 - 3150{,}129 = 43{,}988$$

I kjikvadratfordelinga med 2 fridomsgrader vil ein verdi på 9,21 eller større på observatoren indikere eit signifikansnivå på 1% ($\alpha = 0{,}01$) eller lågare.

c)

Finn formlane for betinga effektplott av samanhengen mellom alder og sannsyn for å kjenne seg svært utrygg når ein er ute og går etter det er mørkt for spanske og norske kvinner som har opplevd kriminalitet i nær familie, har budd i området i 10 år, har 12 års utdanning, er i lønna arbeid og bur i storby utan partnar

Sannsynet for å kjenne seg svært utrygg når ein er ute og går etter det er mørkt er gitt ved

$Pr(Y=1) = 1/(1+\exp\{-E[L_i]\})$ der $E[L_i]$ er forventa verdi av logiten for person "i".

For oppgåve 2 har vi estimert følgjande uttrykk for logiten

$L_i = -0{,}233 +0{,}480*victim_i -0{,}004*yrlvdae_i -0{,}165*eduyrs_i +1{,}519*female_i -0{,}159*liveWithPartner_i +0{,}336*selfempl_i +0{,}430*notempl_i -0{,}028*age_i +0{,}0004*age2_i -0{,}525*suburb_i -0{,}622*town_i -1{,}554*village_i -1{,}455*countryside_i -2{,}274*Spain_i -1{,}722*Sweden_i -1{,}874*Norway_i +0{,}131*eduInSpain_i +0{,}021*eduInSweden_i +0{,}017*eduInNorway_i$

Vi skal finne eit betinga effektplott for samanhengen mellom alder og sannsyn for spanske og norske kvinner som har opplevd kriminalitet i nær familie, har budd i området i 10 år, har 12 års utdanning, er i lønna arbeid og bur i storby utan partnar. Vi må da setje inn verdiar for dette i likninga for logiten.

	Spain	Norway
victim	1	1
yrlvdae	10	10
eduyrs	12	12
female	1	1
liveWithPartner	0	0
selfempl	0	0
notempl	0	0
age	-	-
age2	-	-
suburb	0	0
town	0	0
village	0	0
countryside	0	0
Spain	1	0
Sweden	0	0
Norway	0	1
eduInSpain	12*1	12*0
eduInSweden	12*0	12*0
eduInNorway	12*0	12*1

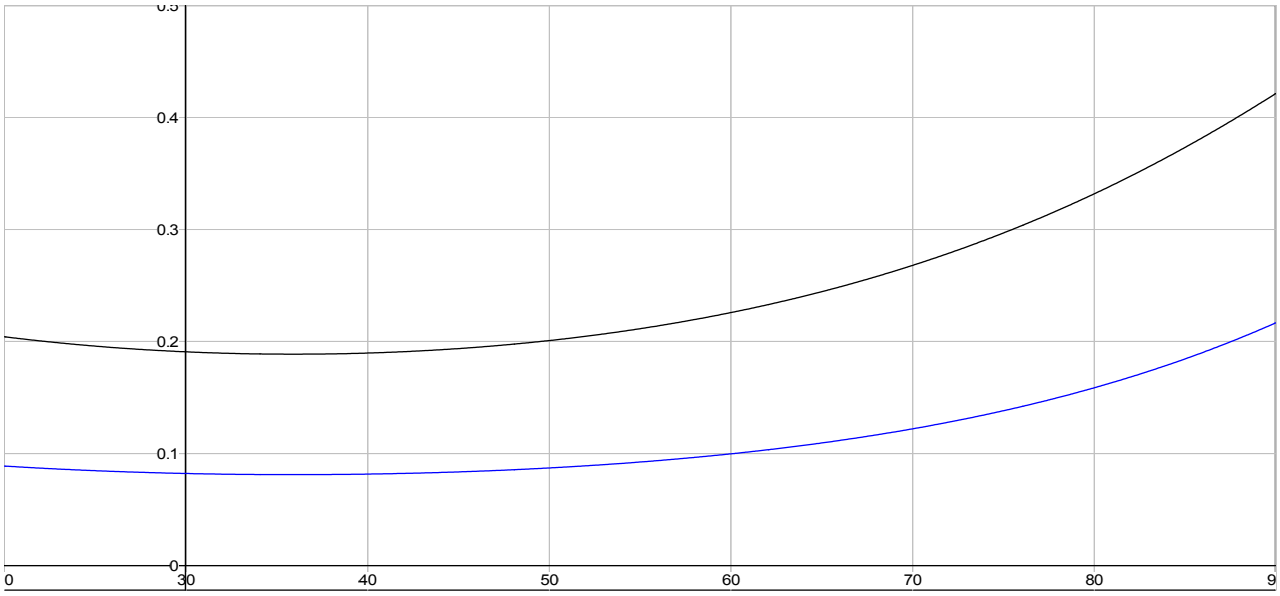
Set vi inn desse variabelverdiane i likninga for logiten forenklar den seg mye
L_i = -0,233 +0,480*1 -0,004*10 -0,165*12+1,519*1 -0,028*age_i +0,00039*$age2_i$ -2,274*$Spain_i$ -1,874*$Norway_i$ +0,131*12*$Spain_i$ +0,017*12*$Norway_i$ =

-0,233 +0,480 -0,04 -1,98 +1,519 -0,028*age_i +0,00039*$age2_i$ -2,274*$Spain_i$ -1,874*$Norway_i$ +1,572*$Spain_i$ +0,204*$Norway_i$ =

-0,254 -0,028*age_i +0,00039*$age2_i$ -2,274*$Spain_i$ -1,874*$Norway_i$ +1,572*$Spain_i$ +0,204*$Norway_i$

Logiten i det betinga effektplott for norske kvinner blir
L_i = -1,924 -0,028*age_i +0,00039*$age2_i$
Formelen for det betinga effektplottet blir
$Pr(Y=1) = 1/(1+exp\{-(-1{,}924 -0{,}028*age_i +0{,}00039*age2_i)\})$

Logiten i det betinga effektplott for spanske kvinner blir
L_i = -0,956 -0,028*age_i +0,0004*$age2_i$
Formelen for det betinga effektplottet blir
$Pr(Y=1) = 1/(1+exp\{-(-0{,}956 -0{,}028*age_i +0{,}00039*age2_i)\})$



$y=1/(1+\exp(-(-0.956 -0.028\times x +0.00039\times x\times x)))$ Spanske kvinner
$y=1/(1+\exp(-(-1.924 -0.028\times x +0.00039\times x\times x)))$ Norske kvinner

Vi ser av figuren at spanske kvinner i den gruppa som er spesifisert her har høgare sannsyn enn tilsvarande norske kvinner for å seie at dei er svært utrygge når dei er ute og går etter det er mørkt. Vi ser også at i begge landa er eldre (over ca 50) kvinner i aukande grad meir utrygge enn dei som er mellom 20 og 50.

d)
Formuler modellen som er estimert

Når ein modell skal formulerast bør det gjerast greie for tre typar element:

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

Med utgangspunkt i data frå Storbritannia, Spania, Sverige og Norge samla inn i 2002 gjennom "European Social Survey" (ESS) er følgjande variablar definert:

Y	**Veryunsa**	Feels very unsafe after dark
X_1	**victim**	Victim of crime in IP's family
X_2	**yrlvdae**	Years lived in the area
X_3	**eduyrs**	Education in years
X_4	**female**	Female or male
X_5	**liveWithPartner**	Lives with a partner
	Employment status	Referansekategori: employed
X_6	selfempl	Selfemployed
X_7	notempl	Not in paid work
	Age	Included as polynom
X_8	age	Age in years
X_9	age2	Age in years squared
	Domicile	Referansekategori: big city
X_{10}	suburb	Lives in a suburb
X_{11}	town	Lives in a town
X_{12}	village	Lives in a village
X_{13}	countryside	Lives in the countryside
	Country	Referansekategori: Storbritannia
X_{14}	Spain	Spain
X_{15}	Sweden	Sweden
X_{16}	Norway	Norway
	EDU*COUNTRY	Interaksjonsledd utdanning*land
X_{17}	eduInSpain	Education in years * Spain (interaction)
X_{18}	eduInSweden	Education in years * Sweden (interaction)
X_{19}	eduInNorway	Education in years * Norway (interaction)

I alle landa som er med i ESS er det gjort tilfeldige utval frå befolkninga slik at det kan trekkjast konklusjonar om befolkninga i kvart land for seg.

I dei fire landa er det i alt 7816 personar som har vore intervjua. Dei fordeler seg slik på dei fire landa:

Spain	1729
United Kingdom	2052
Norway	2036
Sweden	1999
Total	**7816**

Det manglar opplysningar på mange personar på ulike variablar. Etter listevis utelating av manglande data på ein eller fleire av variablane som er definert ovanfor er det igjen 7501 case som kan nyttast i analysen. Det er ingen haldepunkt for å tru anna enn at fråfallet er reint tilfeldig i høve til svar som er registrert på den avhengige variabelen. Listevis utelating er da ein fullgod prosedyre så lenge ein sit igjen med nok case til å gjennomføre analysane. Det gjer vi i dette høvet.

I populasjonen føreset vi at det er eit logistisk samband mellom sannsynet for å ha verdien 1 på den avhengige variabelen (Y) og dei uavhengige X-variablane. Modellen som er estimert 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 . Denne relasjonen skal gjelde for alle "i" som kan reknast inn i befolkningane i Spania, UK, Sverige og Noreg.

Estimert verdi av logiten er definert ved

$$L_i^* = E\left[L_i\right] = \beta_0 + \sum_{k=1}^{19} \beta_k X_{ki}$$

Parametrane i logiten kan estimerast ved ML-metoden (Maksimum Likelihood metoden). Estimata vil vere forventningsrette, variansminimale og normalfordelte, vi kan nytte sannsynsratetesten, og i store utval vil b_k/ SE_{bk} vere tilnærma normalfordelt dersom følgjande føresetnader kan seiast å gjelde:

- 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
- det er ikkje perfekt diskriminering
- det er stort nok utval

e)
Drøft i kva grad føresetnadene for logistisk regresjon er stetta

I punkt d) ovanfor er føresetnadene oppgitt. Ikkje alle føresetnadene kan testast Det kan ikkje ved testing avgjerast om alle relevante variablar er med eller om alle x-variablane er målt utan feil, eller om datainnsamlinga har sikra at alle observasjonane er uavhengige.

To av føresetnadene vil teste seg sjølve. Dersom det i datamaterialet er perfekt multikollinearitet eller perfekt diskriminering[2] vil ikkje estimering av modellen vere

[2] Perfekt diskriminering er ikkje teke med som føresetnader av Hamilton (1992, jfr. side 225 og 233). Perfekt diskriminering representerer substansielt sett same type problem som perfekt multikollinearitet. Begge typane problem fører til store standardfeil for parameterestimata og dermed svært usikre parameterestimat. I høve til

mogeleg. Det faktum at vi har fått eit estimat av modellen viser dermed at desse føresetnadene er oppfylt.

Dei føresetnadene som kan granskast og som skal kommenterast er om 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) og om irrelevante variablar er inkluderte i modellen. Utvalsstorleiken skal også vurderast.

Utvalsstorleik

Storleiken på utvalet er viktig for å sikre nok variasjon til å estimere partielle effektar. I små utval vil det vere større sjanse for å finne case med stor innverknad, høg grad av multikollinearitet og høg grad av diskriminering. Høg grad av multikollinearitet, sterk grad av diskriminering og case med stor innverknad fører til problem for estimeringa i form av upresise estimat (stor varians, stor standardfeil på parameterestimata). Det er ikkje spurt etter kva grad multikollinearitet og diskriminering fører til problem. Case med mogeleg stor innverknad kjem vi attende til nedanfor.

Kva som er eit stort nok utval er ofte problematisk å avgjere. Det har ikkje berre å gjere med talet på case, men og på korleis den avhengige variabelen er fordelt. Svært skeivfordelte Y-variable kan innehalde for lite variasjon til estimeringa sjølv i rimeleg store utval.

Utvalsstorleiken er i estimatet av modellen som er drøfta her på 7501 (vi fin den som summen av case i klassifikasjonstabellen

			Predicted		
			Feeling very unsafe walking alone after dark		Percentage Correct
			0	1	
Observed	Feeling very unsafe walking alone after dark	0	7011	12	99,8
		1	465	13	2,7
	Overall Percentage				93,6

Dette er eit stort utval, men som sagt så er det fordelinga på den avhengige variabelen som er utgangspunktet for vurderinga. Denne fordeling finn vi også i tabellen ovanfor. 478 personar kjenner seg svært utrygge når dei går aleine etter det er mørkt. 478 personar av i alt 7501 er 6,4%. Det er i og for seg tilstrekkeleg, men i store modellar kan det fort bli tale om høg grad av diskriminering. Ser vi

regresjonsresultatet er det problemet med høg grad av multikollinearitet eller høg grad av diskriminering som må drøftast.

vidare på korleis dei dummykoda variablane fordeler seg ser vi at i 3 grupper: "Norway". "Farm or home in countryside" og "Self-employed" er det mindre enn 50 personar som har Y=1. Også " Country village" har eit lågt tal med 62 case der Y=1. Dette kan gi usikre variansestimat for koeffisientane til desse variablane, og det kan gi stor innverknad til case med kombinasjonar av verdiar på desse variablane.

Irrelevante variablar

Test blocks	**2LogLike-lihood of Block**	Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Block 0	-3556,912	Constant	-,233	,534	,190	1	,663	,792
Block 1	-3545,014	Victim	,480	,112	18,275	1	,000	1,616
Block 2	-3194,117	yrlvdae	-,004	,003	1,518	1	,218	,996
		eduyrs	-,165	,027	36,991	1	,000	,848
		female	1,519	,126	144,972	1	,000	4,570
		liveWithPartner	-,159	,111	2,046	1	,153	,853
Block 3	-3150,129	selfempl	,336	,256	1,713	1	,191	1,399
		notempl	,430	,131	10,804	1	,001	1,537
Block 4	-3129,757	Age	-,028	,016	3,071	1	,080	,973
		age2	,000	,000	6,388	1	,011	1,000
Block 5	-3029,321	suburb	-,525	,168	9,747	1	,002	,592
		Town	-,622	,150	17,172	1	,000	,537
		village	-1,554	,183	72,302	1	,000	,211
		countryside	-1,455	,311	21,892	1	,000	,234
Block 6	-2870,363	Spain	-2,274	,371	37,582	1	,000	,103
		Sweden	-1,722	,510	11,426	1	,001	,179
		Norway	-1,874	,624	9,010	1	,003	,154
Block 7	-2849,161	eduInSpain	,131	,032	16,837	1	,000	1,140
		eduInSweden	,021	,045	,208	1	,648	1,021
		eduInNorway	,017	,052	,107	1	,743	1,017

I tabellen ovanfor ser vi av kolonnen kalla "sig" (= sannsynet for å finne ein så stor eller større Wald observator som den vi har gitt i kolonnen kalla Wald) at fleire variablar kan vere irrelevante for modellen. Held vi oss til eit testnivå på 0,01 vil vi måtte konkludere med at både yrlvdae og liveWithPartner ikkje yter signifikant til å forklare variasjonen i Y. Det kan og reisas spørsmål om alder sidan begge ledda i variabelen har sig-verdi over 0,01. Men sidan det er høg grad av multikollinearitet mellom to slike ledd kan vi ikkje stole på estimata av variansen til parameterestimata. Vi bør sjekke ved hjelp av sannsynsratetesten. Alder er inkludert som tillegg i blokk 4 i modellestimatet. Ved å teste om dette tillegget er si signifikant betring av modellen kan vi avgjere om alder er ein relevant variabel.

Vi ser at 2*Loglikelihooden for blokk 3 er -3150,129 og for blokk 4 er den -3129,757. I testen for blokk 4 er det 10 parametrar (K=10), i testen for blokk 3 er det 2 færre parametrar (H=2). Testobservatoren blir da

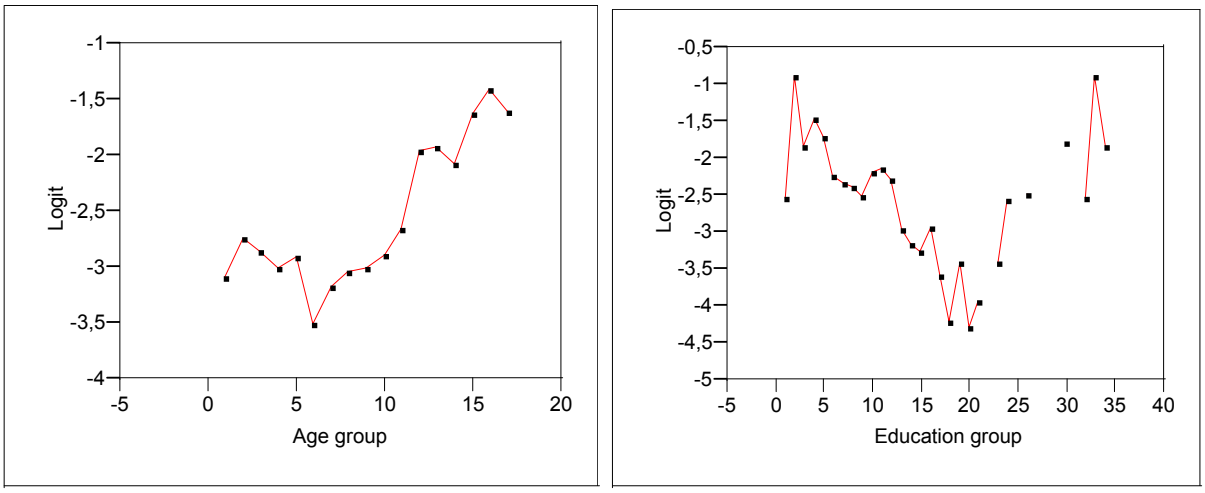
$$\chi^2_2 = -2\{\log_e \mathcal{L}_8 - \log_e \mathcal{L}_{10}\} = -(-3150{,}129 - (-3129{,}757))$$
$$= 3150{,}129 - 3129{,}757 = 20{,}372$$

I kjikvadratfordelinga med 2 fridomsgrader vil ein verdi på 9,21 eller større på observatoren indikere eit signifikansnivå på 1% ($\alpha = 0{,}01$) eller lågare. Alder er altså ein relevant variabel, den yter signifikant til å forklare variasjonen i Y.

To variablar, yrlvdae og liveWithPartner, kunne altså vore droppa frå modellen utan problem.

Det kan her vere verd å merke seg at to av kategoriane i dei dummy koda variablane også har svært låg toleranse. I variabelen sysselsettingsstatus har kategorien "notempl" ein toleranse på 0,026 og i variabelen "Domicile" (bustadstype) har kategorien "countryside" ein toleranse på 0,079. Ei sannsynleg forklaring på dette kan vere at dette er små kategoriar. For "countryside" finn vi 14 personar av 7501 som har svart Y=1. For "notempl" finn vi rett nok 321 personar som her Y=1, men den andre dummyen i variabelen, "self-employed", har berre 20 personar med Y=1. Fordelt på land vil det truleg vere både høg grad av diskriminering og høg grad av multikollinearitet.

<u>Funksjonsform for logiten</u>
Kravet til funksjonell form seier at logiten må vere lineær i parametrane. Dummykoda variable vil alltid vere lineære i denne samanhengen. Sidan vi har konkludert med at yrlvdae (år budd i området) er irrelevant, er det dermed berre utdanning i år og alder i år som kan vere problematiske. Ein studie av kor vidt logiten er lineær for ein gitt variabel kan gjere bruk av at den i så fall må vere lineær for undergrupper av variabelen. Ved å dele opp alder og utdanning i grupper og finne logiten for kvar gruppe kan denne plottast mot grupperinga. Den bør da vere tilnærma ei linje. I tabellvedlegget finn vi slike plott for både utdanning og alder:



For alder ser vi teikn til ein svak kurvesamanheng. For utdanning er diagrammet vanskelegare å tolke. Går vi til histogrammet over utdanning ser vi at det er svært få case med verdiar over 21 år. For fleire kategoriar er det ikkje observasjonar i det heile. Vi kan dermed ikkje legge noko vekt på dei bitane av plottet som er over 21 år. For resten av diagrammet for utdanning er det rimeleg å konkludere med linearitet i logiten.

f)
Vurder om det finst case med uvanleg stor innverknad på regresjonsresultatet

I tabellvedlegget finst boksplott og spreiingsplott for observatorar som kan seie noko om eit case har uvanleg stor innverknad. Dei observatorane vi ser på er særleg ΔPearson kjikvadratet, ΔAvvikskjikvadratet, analogien til Cook's D og leverage. I boksplottet over leverage observatoren er det eitt case som skil seg ut det har identitetsnummer 780. SPSS gir oss den sentrerte leverage verdien slik at ein kan finne leverage ved å legge til gjennomsnittet på K/n (K er talet på parametrar i modellen, n er utvalsstorleiken). I modellestimatet for oppgåve 2 er n=7501 og K=20 slik at gjennomsnittet for leverage vert 0,002666. Leverage for case 780 vert då 0,0475 + 0,0027 = 0,0502. Dette er i alle fall ein liten verdi. Det er dermed neppe noko case som har potensiale for å verke inn på regresjonsresultatet aleine. Dersom det er stor innverknad må det komme på grunn av store residual i prediksjonen. Ser vi på dei to største residualane for ΔPearson kjikvadratet, ΔAvvikskjikvadratet og analogien til Cook's D finn vi fire ulike case. ΔPearson kjikvadratet og ΔAvvikskjikvadratet indikerer dei same to casa, analogen til Cook's D gir to andre case. Desse casa har id.nr. 3135, 202069, 202626 og 110106. Vi finn i vedleggstabellane følgjande data for desse:

Two tables of variable values for 12 cases from the data

IDNO	CNTRY	DOMICIL	CRMVCT	AESFDRK*	YRLVDAE	EDUYRS	EMPL	VICTIM	AGE	FEMALE	LIVEWITH
3135	NO	4	2	4	40	12	1	0	54	0	1
110106	GB	2	1	4	1	29	1	1	57	1	0
202069	SE	4	2	4	0	11	1	0	22	0	1
202626	SE	5	2	4	30	20	3	0	76	1	1

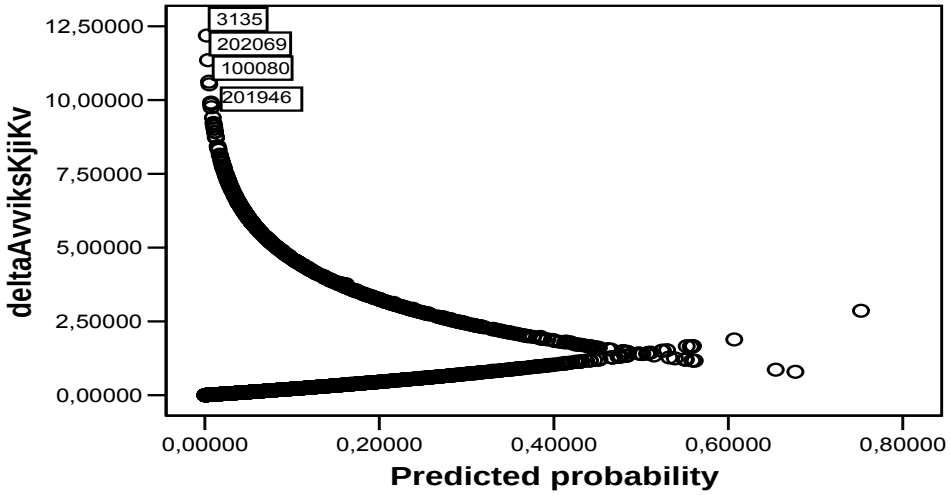
*AESFDRK = 4 tyder at Y=1: desse personane kjenner seg svært utrygge når dei er ute å går aleine etter det er mørkt.

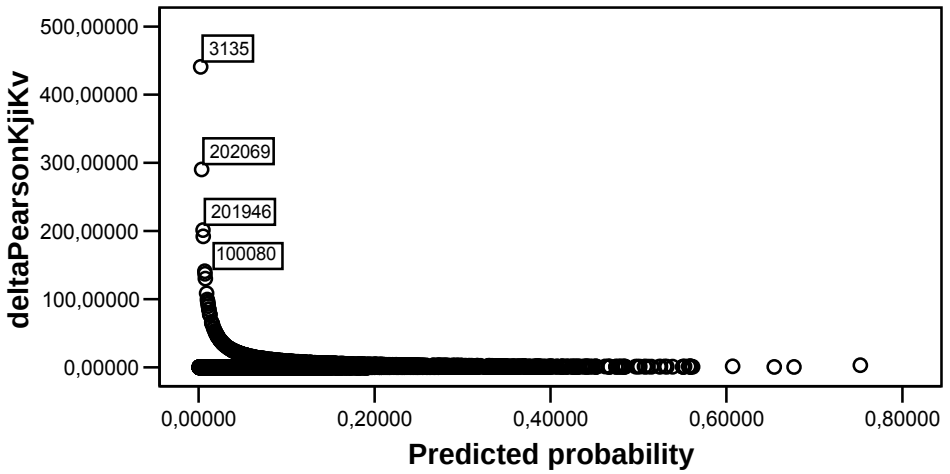
IDNO	PRE	COO	deltaPearsonKjiKv	deltaAvviksKjiKv
3135	0,0023	0,0621	440,915	12,184
110106	0,0211	0,2372	46,715	7,760
202069	0,0034	0,0613	290,272	11,350
202626	0,0116	0,2527	85,211	8,934

Dei eigenskapane som desse personane har gjer at dei får lågt predikert sannsyn. Dermed blir residualen stor. Spørsmålet er om dei er urimeleg store.

Variables	Description	Variables	Description
IDNO	Respondent's identification number	VICTIM	Victim of crime in IP's family
CNTRY	Country	AGE	Age in years
DOMICIL	Domicile, respondent's description	FEMALE	Female respondent
CRMVCT	Respondent or household member victim of burglary/assault last 5 years	LIVEWITH	Lives with partner
AESFDRK	Feeling of safety of walking alone in local area after dark	PRE	Predicted probability
YRLVDAE	How long lived in this area	COO	Analog of Cook's influence statistic
EDUYRS	Years of full-time education completed	deltaPearsonKjiKv	Change in Pearson chi square
EMPL	Employment status	deltaAvviksKjiKv	Change in Deviance chi square

Det er her 4 case som kanskje kan ha uønska stor innverknad på regresjonsresultatet. Verdiane på både ΔPearson kjikvadratet og ΔAvvikskjikvadratet er svært høge. Begge observatorane måler kor dårleg modellen passar for case ”i” og er asymptotisk kjikvadratfordelt (Hamilton 1992: 237). Ein verdi på observatoren over 4 vil dermed indikere at fjerning av caset vil endre modellen signifikant. Vurdert ut frå plotta er det svært mange case dette gjeld for. No skal ein hugse at når n er stor skal det jamt over svært lite til før ei endring er signifikant. Det er i tillegg nødvendig at endringa er substansielt interessant.





Den mest praktiske tilnærminga til spørsmålet om dei faktisk har uønska innverknad vil derfor vere å estimere modellar der ein utelet eitt og eitt av desse fire casa for å sjå om det fører til substansielle endringar i parametrane. Slike modellar er ikkje presentert i oppgåveteksten. Før ein gjennomfører desse estimata bør ein også utelate dei irrelevante variablane som er identifisert ovanfor.

Det kan vere verd å merke seg at case nr 3135 (som har størst innverknad etter ΔPearson kjikvadratet og ΔAvvikskjikvadratet) er norsk og bur i tettstad, medan case 202626 (som har størst innverknad etter analogien til Cook's D) er busett på landet og utan betalt arbeid. Det kan synast som at stor innverknad har samanheng med tendensen til stor grad av diskriminering.