Arbeidsnotat nr. 8-2000

Per Arne Tufte

En intuitiv innføring i logistisk regresjon

SIFO

Prosjektnotat nr. 8 - 2000

STATENS INSTITUTT FOR FORBRUKSFORSKNING
Sandakerveien 24 C, Bygg B
Postboks 4682 Nydalen
0405 Oslo
www.sifo.no

En intuitiv innføring i logistisk regresjon

Av

Per Arne Tufte

JUNI 2000
STATENS INSTITUTT FOR FORBRUKSFORSKNING
LYSAKER

Forord

Dette notatet har som hensikt å gi en så elementær innføring i prinsippene for logistisk regresjon som mulig. Notatet bygger på et upublisert manuskript fra 1999. Det utgis i SIFOs notatserie fordi det har vist seg å være en del etterspørsel etter en elementær innføring i denne metoden.

En rekke kolleger på SIFO har lest og kommentert tidligere utkast til dette notatet. Jeg vil spesielt takke Christian Poppe, Johan Håkon Bjørngård, Randi Lavik, Elin Bye, Tove Mordal og Margareta Wandel for nyttige kommentarer og innspill. De er naturligvis ikke ansvarlige for i det foreliggende notatet. Forfatteren mottar gjerne ytterligere kommentarer til det foreliggende notatet

JUNI 2000

STATENS INSTITUTT FOR FORBRUKSFORSKNING

Innhold

1 Innledning

Regresjonsanalyse har i løpet av det siste tiåret blitt den dominerende kvantitative analyseformen innen samfunnsforskningen. En viktig årsak til dette er metodens fortrinn når det gjelder multivariate analyser, dvs. analyser hvor flere uavhengige variabler trekkes inn for å belyse en avhengig variabel. I slike analyser kommer raskt tradisjonell krysstabellanalyse til kort.

Det kunne sies mye om den faktiske anvendelse av regresjonsanalyse. Satt på spissen synes det som om regresjonsanalyse betraktes som et universalverktøy til bruk for nær sagt enhver problemstilling. Målenivået på den avhengige variabel problematiseres ikke, R^2 godtas ukritisk som et egnet mål på modellens tilpasning til dataene, det er lite bekymring om hvorvidt forutsetningene for regresjon er tilfredsstilt, residualanalyser og annen diagnosestatistikk ignoreres. I tillegg er de modeller som testes ut i liten grad teoretisk fundert. Det godtas ofte ukritisk at sammenhenger er additive og lineære, enda teorien kanskje tilsier noe annet. Byggingen av regresjonsmodeller har preg av et ritual hvor det fylles på med mer eller mindre relevante uavhengige variabler for å «kontrollere» for disse. Men hva hjelper det for eksempel å kontrollere for en kontinuerlig aldersvariabel, dersom forskjellene i virkeligheten er knyttet til alderskategorier?

Et annet moment er hensikten med undersøkelsen. Det er forskjell på en analyse hvor siktemålet er å undersøke effekten av en uavhengig på den avhengige (faktororientering) og en analyse hvor en tilstreber en teste ut hvor god en modell er når det gjelder å forklare variasjon i den avhengige variabelen (modellorientering). I det første tilfellet er det viktig å inkludere alle variabler som både påvirker den avhengige og den uavhengige variabelen for å unngå spuriøse sammenhenger. I det andre tilfellet er det interessant å inkludere alle kilder til variasjon i den avhengige variabelen (Hagquist & Stenbeck 1998:230-31). Et tredje siktemål kunne være å sammenlikne grupper som er mest mulig like med hensyn til ulike egenskaper (variabler), ikke for å teste kausalhypoteser, men for eksempel for å avdekke sentrale sosiale skillelinjer.

Svært mye av grunnen til denne ukritiske bruk av regresjonsanalyse ligger i metodebøkenes fokusering på prinsippene bak regresjon, og ikke på hvilke konsekvenser det har når forutsetningene er brutt (hva er f.eks. konsekvensene har det f.eks. når residualene er heteroskedastiske?). I tillegg legges det ofte stor vekt på at regresjonsanalyse er en robust teknikk og at konsekvensene av bristende forutsetninger som oftest er relativt ubetydelige. Holdningen til regresjonsanalysens evne til å absorbere slike brudd er nærmest «tar du den ,så tar du den». Selv om enkelte brister ikke nødvendigvis får store følger, kan produktet av flere brister samtidig bli uforutsigbart. Enkelte forutsetninger er også mer kritiske enn andre.

I dette notatet tar jeg opp en meget vesentlig forutsetning for ordinær regresjon, nemlig at den avhengige variabelen i prinsippet må være metrisk eller kvantitativ, dvs. enten på intervall- eller forholdstallsnivå. Kvalitative variabler, dvs. variabler på nominal- eller ordinalnivå eller variabler med to verdier (dikotome variabler) tilfredsstiller ikke forutsetningene for regresjon basert på minste kvadraters metode. For variabler på nominalnivå skyldes dette at tallverdiene på variablene ikke gjenspeiler noen rangering. Verdiene kan endres vilkårlig uten at dette har noen som helst substansiell betydning. For variabler på ordinalnivå kan derimot verdiene rangeres. Problemet her er imidlertid avstanden mellom kategoriene. Er f.eks. avstanden mellom å være helt enig og litt enig i en påstand, like stor som avstanden mellom å være helt uenig og litt uenig? Det at regresjonsanalyse i prinsippet ikke kan behandle slike variabler, betyr imidlertid ikke at analysen bryter sammen når avhengig variabel er kvalitativ. Likningene, matrisene og algoritmene bak analyseresultatene ser nemlig bare tall og vil som regel komme fram til en løsning uansett hvilke målenivå variablene har. Derfor er det viktig at forskeren er klar over de begrensninger som ligger i metoden.

En metode for å behandle kvalitative, avhengige variabler er logistisk regresjon. Metoden har tiltatt i popularitet de siste årene. Fra å være relativt lite brukt på begynnelsen av 90-tallet, er den i dag nesten

den dominerende formen for regresjonsanalyse innen sosiologisk og statsvitenskapelig forskning. En av de viktigste grunnene til dette er naturligvis at langt de fleste avhengige variabler disse fagdisiplinene arbeider med er dikotome eller på nomial-/ordinalnivå.

På en måte er denne utviklingen uheldig. På et tidspunkt hvor det fortsatt er sviktende kunnskaper om den enkleste formen for regresjonsanalyse, overtar en ny teknikk som er enda vanskeligere tilgjengelig enn den forrige. I motsetning til OLS (ordinær regresjon ved bruk av minste kvadraters metode) er ikke resultatene fra logistisk regresjon intuitivt lett tilgjengelig. Dette gjelder spesielt de estimerte koeffisientene som ikke har noen enkel og lett forståelig fortolkning.

Matematikken bak logistisk regresjon og annen form for Maximum Likelihood regresjon er ikke lett tilgjengelig. Modellene kan i praksis ikke løses gjennom algebra, men må beregnes ved hjelp av en iterativ algoritme - kall det en prøve-og-feile prosess, selv om framgangsmåten er langt mindre tilfeldig enn som så. Som oftest skal det ikke så mange interasjonene til før man finner en løsning. For dem som ikke forstår matematikken bak «minste kvadraters metode» er håpet om å forstå matematikken bak logistisk regresjon enda mindre.

Samtidig er det etter min mening viktig at man vet hva man driver med, dvs. at man har en grunnleggende forståelse av hvordan logistisk regresjon virker. Hensikten med denne innføring er å gi en innføring i logistisk regresjon, hvor matematikken er redusert til et minimum. Jeg har i størst mulig grad forsøkt å gi en intuitiv innføring i metoden. Forbildet er Gudmund Hernes' utmerkede artikkel «En intuitiv innføring i multivariat analyse» (1976:147-188), men resultatet i form av dette notatet står selvfølgelig langt tilbake for dette.

Planen for notatet er å først ta utgangspunkt i en analyse hvor den avhengige variabelen har to verdier og utvikle prinsippet for logistisk regresjon ut fra dette. Deretter utvides den enkle modellen, til først å behandle variabler på nominalnivå med mer enn to verdier, deretter variabler på ordinalnivå med mer enn to verdier. Framstillingen forutsetter at leseren er kjent med krysstabellanalyse og multippel regresjonsanalyse. Skog (1998) gir en grunnleggende og meget god innføring på norsk i kausalanalyse og regresjonsanalyse.

En advarsel til slutt. Selv om logistisk regresjon kan behandle kvalitative avhengige variabler, betyr ikke det at denne teknikken er et nytt universalverktøy for enhver anledning. Det er spesielt viktig å tenke igjennom på forhånd om det er rimelig å tenke seg at den avhengige variabelen oppfører seg slik modellen forutsetter. Det finnes alternativer til logistisk regresjon som kan være bedre egnet til enkelte problemstillinger.

2 Regresjon med dikotom avhengig variabel

2.1 Kategoriske variabler

Det kan innledningsvis være nyttig å avklare tre sentrale begreper for å karakterisere et datamateriale som skal analyseres eller problemstillinger i en undersøkelse, nemlig enheter, variabler og verdier (Galtung 1970). *Enhetene* er dem vi ønsker å få vite noe om eller dem som et datamateriale gir informasjon om. I samfunnsvitenskapelige undersøkelser er enhetene som oftest individer, men de kan også være f.eks. grupper, organisasjoner eller nasjoner. *Variablene* dreier seg om hva det er vi vil vite om enhetene i undersøkelsene, dvs. egenskaper som beskriver noe ved enhetene. Dette kan være egenskaper fra kjønn og alder til holdninger og konkrete handlinger. *Verdiene* på variablene gir informasjon om hvordan de konkrete egenskapene vi måler er gruppert. Variabelen alder kan f.eks. være inndelt i aldersgrupper, variabelen kjønn i verdiene mann og kvinne.

I metodelitteraturen opererer en med ulike inndelinger av variabler i kategorier. En type inndeling tar utgangspunkt i hvilket nivå den egenskapen variabelen sier noe om befinner seg på i forhold til undersøkelsesenheten. En kan her blant annet snakke om absolutte variabler dersom egenskapen gjelder undersøkelsesenheten selv og kontekstuelle variabler dersom egenskapen gjelder systemer som enheten er medlem av (familie, nærmiljø, lokalsamfunn, organisasjoner etc.). En annen type inndeling sier noe om hva slags egenskaper som måles. Ved individundersøkelser skilles det ofte mellom bakgrunnsvariabler, personlighetsvariabler og holdnings- og atferdsvariabler.

I vår sammenheng er det to andre inndelinger som er av interesse. Den første skiller mellom variabler utfra hvordan de er inndelt i verdier. På den ene siden har vi kontinuerlige variabler som kan inndeles i et uendelig antall verdier, avhengig av hvor nøyaktig vi ønsker å måle egenskapen. Alder kan f.eks. deles inn i år, måneder, uker, dager, timer, minutter, sekunder, etc. På den annen side har vi diskrete variabler som har et avgrenset antall verdier.

Den andre inndelingen skiller mellom målenivå på variablene og har paralleller til skillet mellom kontinuerlige og diskrete variabler. Det opereres gjerne med fire målnivåer: nominalnivå, ordinalnivå, intervallnivå og forholdstallsnivå. For variabler på nominalnivå angir verdiene kun at enhetene er like eller ulike med hensyn til den egenskapen variabelen måler. Verdiene angir ikke noen som helst rangering på variabelen. Nasjonalitet eller bosted kan være eksempler på slike variabler. På det neste målenivået, ordinalnivå, angir verdiene en rangering, men det er mulig å si noe eksakt om avstanden mellom verdiene. En holdningsspørsmål med verdiene svært enig, meget enig, litt enig, litt uenig, meget uenig og svært uenig, kan være et eksempel på en slik variabel. For de to høyeste målnivåene, intervallnivå og forholdstallsnivå opereres det med en eksakt avstand mellom verdiene på en variabel. Forskjellen mellom disse målenivåene er at det sistnevnte forutsetter et absolutt nullpunkt, noe det første ikke gjør. Dette skillet er som oftest av mindre interesse innenfor samfunnsvitenskapene. Eksempler på variabler på forholdstallsnivå er alder, inntekt og gjeld.

Kontinuerlige variabler er alltid på forholdstallsnivå, men det motsatte er ikke nødvendigvis tilfelle. Ta for eksempel en variabelen antall barn i husholdningen. Variabelens er på forholdstallsnivå, men verdiene kan samtidig ikke deles opp i det uendelige. Det gir ikke mening å snakke om 0,5 barn eller liknende. Ordinær regresjonsanalyse forutsetter strengt tatt at den avhengige variabelen er kontinuerlig, dvs. at den kan inndeles i et uendelig antall verdier. I praksis er det ikke noe problem å benytte diskrete variabler på forholdstallsnivå, men fortolkningen av koeffisientene kan da ofte bli noe merkelig.

Variabler på nominalnivå og ordinalnivå kalles ofte for kvalitative, ikke-metriske eller kategoriske variabler (se f.eks. Hellevik 1991:155, Agresti 1997:1-2). Slike variabler byr på problemer når de skal benyttes i en regresjonsanalyse. Brukt som uavhengige variabler kan de kodes om til dummyvariabler (se nedenfor), men brukt som avhengige variabler bryter de med forutsetningene for ordinær regre-

sjonsanalyse. En måte å løse disse problemene på er å benytte ulike varianter av logistisk regresjon. Før vi går inn på logistisk regresjon med variabler på nominal- og ordinalnivå, tar vi imidlertid utgangspunkt i regresjonsanalyse med en dikotom avhengig variabel.

2.2 Dikotome variabler – sannsynligheter og andeler

Teknisk sett er en dikotom variabel en variabel med to mulige verdier. Andre betegnelser på slike variabler er binære variabler eller dummyvariabler (Greene 1993:229). Verdiene på variabelen antyder hvorvidt enhetene i undersøkelsen har en bestemt egenskap eller ikke. Som regel gis verdien 0 dersom egenskapen ikke er tilstede og verdien 1 dersom egenskapen er tilstede. Slike egenskaper kan være handlinger, kunnskaper, holdninger, kjennetegn eller andre fenomener. Eksempler på dikotome variabler kan være:

- handlinger/beslutninger	-	kjøpt svarte tjenester (1) eller ikke (0)
- kunnskaper	-	hørt om kjøpsloven (1) eller ikke (0)
- holdninger	-	tilhenger av selvbestemt abort (1) eller ikke (0)
- sosiale kjennetegn	-	har høyere utdanning (1) eller ikke (0)
- andre fenomener	-	opplevd å bli arbeidsledig siste år (1)eller ikke (0)

Slik disse variablene er definert, gir det ikke mening å snakke om verdier som ligger midt mellom 0 og 1. Enten er egenskapen, handlingen, kunnskapen etc. tilstede, eller så den det ikke. Det er også meningsløst å snakke om verdier som er høyere enn 1 eller lavere enn 0. Det er med andre ord et kvalitativt sprang mellom 0 og 1. Slike variabler kalles derfor også ofte for kvalitative variabler.

Dikotome variabler, både som avhengige og uavhengige, er spesielt egnet i krysstabellanalyser, fordi de forenkler analysearbeidet. I en bivariat krysstabellanalyse med dikotom avhengig variabel kan en sammenlikne andelen som har et kjennetegn (verdien 1 på den avhengige variabelen) mellom grupper av enheter med ulike verdier på den uavhengige variabelen. Forskjeller i andeler kan uttrykkes som prosendifferanser, men kan også fortolkes som forskjeller i sannsynligheter for å ha verdien 1 på den avhengige variabelen.

Måten dikotome, avhengige variabler behandles på i regresjonsanalyse likner måten de håndteres på i krysstabellanalyse. Regresjonsanalyse, også logistisk, behandler som nevnt alle variabler som om de er kontinuerlige. Den dikotome variabelen må derfor fortolkes som en representasjon av en «underliggende» kontinuerlig variabel, dvs. en variabel som i prinsippet kan anta alle mulig verdier i intervallet 0 og 1. Det innebærer at de resultater vi kommer fram til ikke kan fortolkes direkte i tilknytning til verdiene 0 og 1 på den avhengige variabelen. I stedet må resultatene fortolkes som forskjeller i andeler eller sannsynligheter for å ha verdiene 0 eller 1 på den avhengige variabelen, akkurat slik det gjøres i krysstabellanalysen. Det er viktig å se denne parallellen mellom krysstabellanalyse og regresjonsanalyse når den avhengige variabelen har to verdier.

Vi kan følgelig si at den dikotome avhengige variabelen må omformes eller omfortolkes. Denne omformingen kan illustreres på to parallelle måter:

2.2.1 Omforming til en underliggende sannsynlighet/tilbøyelighet

Vi tenker oss at det er en underliggende, kontinuerlig sannsynlighet eller tilbøyelighet for at egenskapen eller fenomenet skal inntreffe og at denne underliggende sannsynligheten påvirkes av eller fordeler seg ulikt i forhold til en del uavhengige variabler. Hver enkelt enhet i undersøkelsen påvirkes av denne underliggende sannsynligheten.

Sannsynligheter
Sannsynlighetsbegrepet kan illustreres ved hjelp av myntkast. Dersom man kaster en ordinær mynt, skulle det normalt være 50/50 prosent sjanse for at kastet gir enten krone eller mynt. Det innebærer at en av hundre kast skulle en forvente å få 50 kast med krone og 50 med mynt. Dette forholdet kan uttrykkes ved å si at sannsynligheten for henholdsvis mynt eller krone er 0,5. Dersom vi opererte med en falsk mynt hvor kronesiden var trykt på begge sider, ville vi bare få krone og ingen mynt. Da ville sannsynligheten for å få krone være lik 1 og sannsynligheten for å få mynt være lik 0. Vi ville med andre være sikre på å få krone og ikke mynt. Dette illustrerer at sannsynligheter alltid befinner seg i intervallet fra 0 til 1. Er sannsynligheten lik 0, er det sikkert at et fenomen ikke inntreffer. Er den 1, er det sikkert at det inntreffer. Ligger sannsynligheten mellom 0 og 1 angir det forskjellige grader av usikkerhet knyttet til om fenomenet inntreffer eller ikke.

Den observerte fordelingen på den dikotome avhengige variabelen gjenspeiler hvordan den underliggende sannsynlighetsfordelingen har slått ut blant respondentene i undersøkelsen. Når sannsynligheten for at fenomenet inntreffer er under 0,5 er det mest sannsynlig at fenomenet ikke inntreffer. Når sannsynligheten for at fenomenet inntreffer er over 0,5 er det mer sannsynlig at fenomenet inntreffer enn at det ikke inntreffer. Er sannsynligheten lik 0,5 er det like sannsynlig at fenomenet inntreffer som at det ikke gjør det.

Dersom vi har en dikotom variabel som definerer en egenskap, et fenomen e.l., kan sannsynligheten for at fenomenet eller egenskapen forekommer hos enheter i en gruppe enkelt beregnes ut ved å dele antall enheter som har verdien 1 med det totale antall enheter i gruppen.

Eksempel:
Vi sammenlikner 10 kvinner med 20 menn når det gjelder hvorvidt de har kjøpt svarte tjenester eller ikke i løpet av det siste året. Mens 2 av kvinnene oppgir at de har kjøpt svart, gjelder dette 8 av mennene. Vi kan ut fra disse opplysningene beregne sannsynligheter for å ha kjøpt svart i begge gruppene:

Sannsynligheten for å ha kjøpt svart blant kvinnene er 2/10 = 0,2
Sannsynligheten for å ha kjøpt svart blant mennene er 8/20 = 0,4

Sannsynligheten for å ha kjøpt svarte tjenester er med andre ord dobbelt så høy blant mennene som blant kvinnene i utvalget.

2.2.2 Omforming til andeler

En annen betraktningsmåte, som ikke står i motsetning til den første, er å tenke seg at fenomenet har en viss utbredelse både totalt og i forskjellige grupper av enheter. Avhengig av hvilken gruppe en undersøker, er det en bestemt andel av enhetene som har egenskapen, er utsatt for fenomenet, har en bestemt holdninger, har handlet på en bestemt måte etc. Mens begrepet sannsynlighet beskriver en tilbøyelighet eller tendens hos det enkelte individ i undersøkelsen til å ha en egenskap, være eksponert for et fenomen e.l., viser andeler til utbredelsen av egenskapen/fenomenet i grupper av enheter.

Den observerte fordelingen på den dikotome avhengige variabelen danner rett og slett utgangspunkt for å regne ut andelen som har verdien 1 på den avhengige variabelen. Vi ser følgelig at utregningen av andeler er identisk med og gir samme resultat som beregningen av sannsynligheter. Andeler og sannsynligheter er her parallelle begreper.[1]

[1] Skillet mellom sannsynligheter og andeler kan virke kunstig. En kan f.eks. tenke seg sannsynlighetsbegrepet anvendt på gruppenivå. Poenget her er å skissere en alternativ tankegang til den individualistiske sannsynlighetstekningen som etter min mening er problematisk når en skal fortolke resultater blant annet i forbindelse med logistisk regresjon.

Eksempel:
Vi kan følge opp eksempelet med kjøp av svarte tjenester. En sannsynlighet på 0,2 innebærer rett og slett at 20 prosent av kvinnene har kjøpt svart, mens en sannsynlighet på 0,4 innebærer at 40 prosent av mennene har kjøpt svart. Som vi så ved sammenlikning ved sannsynlighetene, er andelen som har kjøpt svart dobbelt så blant mennene enn blant kvinnene.

Begge betraktningsmåtene ovenfor er riktige, men utfyller samtidig hverandre. Det er viktig å ha i mente at en underliggende sannsynlighet eller tilbøyelighet på individnivå gir seg utslag i bestemte andeler på gruppenivå. Måten vi regner oss fram til sannsynlighetene på er via andeler av enhetene i disse gruppene som har verdien 1 på variabelen. Det er også viktig å se parallellen til krysstabellanalyse. En krysstabell basert på eksempelet ovenfor ville sett som følger:

Tabell 2.1: Antall som har kjøpt svarte tjenester avhengig av kjønn i et tenkt eksempel. Antall og prosent.

Kjønn	***Mann (0)***	***Kvinne (1)***	**Totalt**
Kjøpt svarte tjenester siste år?			
Nei (0)	60 % (12)	80 % (8)	67 % (20)
Ja (1)	40 % (8)	20 % (2)	33 % (10)
Sum	100 % (20)	100 % (10)	100 % (30)

Her finner vi igjen andelene uttrykt i prosent. Som vi har vist kan disse raskt gjøres om til sannsynligheter (20% andel er lik en sannsynlig på 0,2). Prosentdifferansene uttrykker som tidligere nevnt forskjeller mellom andeler som har f.eks. verdien 1 på den avhengige variabelen.

2.3 Ordinær regresjon og sannsynligheter – lineær sannsynlighetsregresjon

Beregningen av sannsynligheter eller andeler på bakgrunn av dikotome variabler er første skritt i retning av å vise hvordan slike variabler kan håndteres som avhengige variabler i regresjon. Men ett spørsmål gjenstår: I motsetning til krysstabellanalyse, hvor enheter grupperes i celler (kombinasjoner av verdier på variablene) og telles opp, arbeider ikke regresjonsanalyse med grupper av enheter, men individuelle data. Hvordan beregner så regresjonsanalysen sannsynligheter for utfallet av en dikotom variabel når det eneste vi har oppgitt av verdier på denne variabelen er 0 og 1 for hver enkelt observasjon.

Svaret er faktisk at dette automatisk blir resultatet ut fra beregningsmåten ved ordinær regresjon. Hvorfor det blir slik kan det i første omgang være vanskelig å begripe. Det er imidlertid lettere å forstå dette når vi har i mente at koeffisientene i regresjonsanalyse alltid viser endringer i gjennomsnittsverdien på den avhengige variabelen når den uavhengige øker med en enhet i verdi. La oss ta et enkelt eksempel på en bivariat regresjon med timelønn som avhengig variabel og kjønn som uavhengig. Dersom den ustandardiserte regresjonskoeffisienten for kjønn er - 10 kroner, innebærer det at kvinner i gjennomsnitt tjener 10 mindre i timen enn menn.

Hva sier gjennomsnittsverdien for en dikotom variabel med verdiene 0 og 1? Gjennomsnittet finner vi ved å summere alle verdiene som enhetene i et utvalg har på variabelen, og så dele på antall enheter. Summen av en variabel med verdiene 0 og 1 vil rett og slett være antall enheter som har verdien 1. Gjennomsnittet, som vi får ved å dele denne summen på totalt antall enheter, vil følgelig være andelen som har verdien 1 i utvalget.

Eksempel:
Gjennomsnittet av variabelen kjøpt svart arbeid for kvinner finner vi ved å summere de ti observasjonene og dele på totalt antall observasjoner:

(0+0+0+0+0+0+0+0+1+1)/10 = 2/10 = 0,2

Dette tallet er identisk med sannsynligheten eller andelen som vi beregnet tidligere. Hvis vi tar gjennomsnittet blant et utvalg enheter av en dikotom variabel med verdiene 0 og 1, vil resultatet alltid være andelen i dette utvalget som har verdien 1 på variabelen.

Det at de estimerte koeffisientene i regresjonsanalyse viser gjennomsnittsendringer i avhengig variabel ved å øke verdien på en uavhengig variabel med en enhet, innebærer altså at koeffisientene, når vi har en dikotom avhengig variabel kodet med verdiene 0 og 1, kan fortolkes som endringer i andeler eller sannsynligheter for å ha verdien 1 på den avhengige variabelen som følge av en enhets endring i en uavhengig variabel. Følgelig ser det i først omgang ikke ut til å være noe i veien for å benytte ordinær lineær regresjonsanalyse når vi har en dikotom avhengig variabel. Denne analyseformen kalles for *lineær sannsynlighetsregresjon*.

2.4 Alt såre vel? – svakheter ved lineær sannsynlighetsregresjon

Så langt virker alt vel. Dersom lineær regresjon tross alt kan behandle dikotome variabler, skulle man kunne nøye seg med denne analyseformen og grunnlaget for dette notatet ville være relativt tynt. Imidlertid er lineær sannsynlighetsregresjon beheftet med en del svakheter i form av brudd på forutsetningene for ordinær regresjon. Disse svakhetene er at feilleddene ikke er normalfordelte, at feilleddene er heteroskedastisk fordelt, at modellen kan gi meningsløse prediksjoner og at den funksjonelle formen på sammenhengen ofte er tvilsom.

Den første svakheten, at feilleddene ikke er normalfordelte, er ikke den mest vesentlige. Dette vil ikke skape problemer for estimatene av koeffisienter, slik at disse fortsatt er forventningsrette. Derimot har denne forutsetningen betydning for å kunne trekke nøyaktige statiske slutninger på bakgrunn av resultatene. Når det gjelder de andre punktene ovenfor vil jeg drøfte hvert av dem nærmere nedenfor.

2.4.1 Heteroskedastisitet

Lineær regresjon forutsetter at feilleddene er fordelt homoskedastisk. Det innebærer at spredningen rundt regresjonslinjen, dvs. avstandene mellom predikerte og observerte verdier, er omtrent den samme uavhengig av hvilke verdier vi har på den uavhengige variablene. Dersom variasjonen rundt regresjonslinjen ikke er konstant, har vi heteroskedastisitet. Ved regresjon med dikotom avhengig variabel vil feilleddene være heteroskedastiske fordi den betingede variansen til feilleddene er avhengig av de uavhengige variablene og følgelig ikke konstant.[2]

Heteroskedastiske feilledd virker ikke inn på parameterestimatene (regresjonskoeffisientene). Disse vil fortsatt være forventningsrette. Derimot vil ikke lenger standardfeilen og de mål som bygger på denne

[2] En matematisk utlegning av dette finnes blant annet i Long (1997:38).

– konfidensintervaller, t-tester og F-tester – være korrekte. Det er heller ikke mulig å si noe om standardfeilene er systematisk for høye eller for lave (Gujarati 1988:325-26). Vi kan følgelig ikke trekke slutninger om at signifikanstestene ved heteroskedastisitet innebærer økt sannsynlighet for feil av type I (beholde en gal nullhypotese) eller type II (forkaste en riktig nullhypotese).

Heteroskedastisitet er et alvorlig problem, men problemet har vært foreslått løst ved hjelp en totrinnsprosedyre hvor en i første omgang estimerer predikerte sannsynligheter med vanlig regresjon og i neste omgang estimerer koeffisientene ved hjelp av veide minste kvadraters metode («WLS).[3] Selv om denne framgangsmåten løser problemet med heteroskedastisitet og øker effektiviteten knyttet til estimatene, løser det ikke andre og mer grunnleggende problemer med sannsynlighetsregresjon.

2.4.2 Problemet med å begrense avhengig variabel til å ligge mellom 0 og 1

Som nevnt må sannsynligheter (eller andeler) befinne seg i intervallet 0 og 1. Det gir ikke mening å snakke om sannsynligheter som er lavere enn 0 eller høyere enn 1. Ved lineær sannsynlighetsregresjon vil en imidlertid ikke sjelden predikere slike meningsløse sannsynligheter, noe som også utgjør den vesentligste svakheten ved lineær sannsynlighetsregresjon. Vi kan illustrere dette ved hjelp av et eksempel.

Eksempel:
Vi foretar en lineær sannsynlighetsregresjon med timelønn som avhengige variabel og kjønn og antall år med utdanning utover grunnskole som uavhengige. Lønn er en dikotomisert variabel som er forholdsvis skjevfordelt. Verdien "høy lønn" er de som tjener kr. 140 eller mer i timen. Dette utgjør 9,2 prosent av enhetene i undersøkelsen.

Tabell 2.2: Lønnsnivå (dikotomiert) avhengig av kjønn og utdanning. Arbeidslivsundersøkelsen 1993. Utskrift fra lineær sannsynlighetsregresjon.

Coefficients

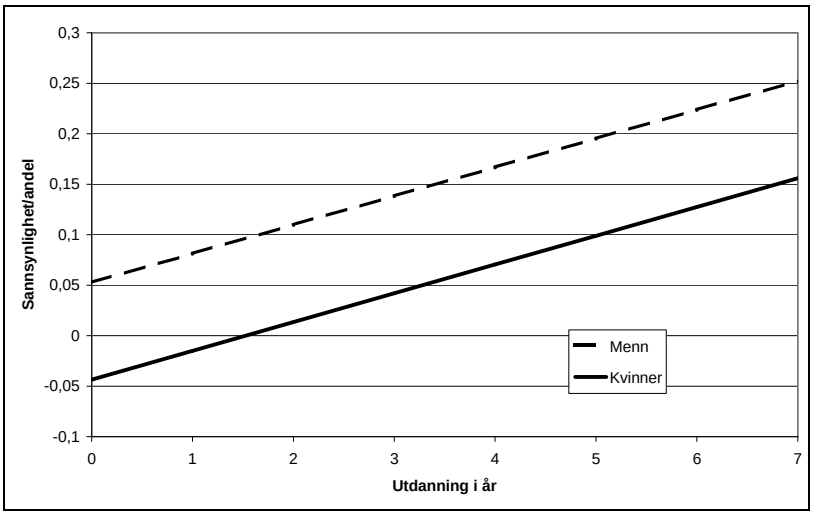
		Unstandardized Coeffisients		Standardized Coeffisients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	0,0528	0,0088		5,9815	0,0000
	kjønn	-0,0963	0,0097	-0,1679	-9,8846	0,0000
	utdanning i år	0,0285	0,0019	0,2568	15,1174	0,0000

a Dependent Variable: lønn dikotomisert

Resultatene fra regresjonsanalysen er vist i tabellen ovenfor. Konstanten i utskriften viser predikert sannsynlighet for å ha høy lønn for menn som ikke har utdanning utover grunnskole (verdien 0 på begge de uavhengige variablene). Denne sannsynligheten er 0,05 og kan også tolkes som andelen som har høy lønn i denne gruppen (5 prosent). Den negative kjønnskoeffisienten viser at sannsynligheten for å ha høy lønn i gjennomsnitt er 0,09 lavere blant kvinner enn menn, kontrollert for utdanning. Utdanningskoeffisienten viser at sannsynligheten for å ha høy lønn i gjennomsnitt er 0,03 høyere for hvert år mer med utdanning respondentene i undersøkelsen har, kontrollert for kjønn.

[3] Interesserte i WLS og den mer generelle GLM («generalized least squares method») henvises til lærebøker i regresjon og økonometri, f.eks. Greene (1993), Guijarati (1988:321-25 og 337-38), Aldrich and Nelson (1984).

Regresjonsresultatet kan også illustreres ved en figur, slik som nedenfor. Figuren viser, som vi allerede har kommentert, at kvinner generelt har en lavere sannsynlighet enn menn for å ha høyere lønn, men at utdanning øker denne sannsynligheten både for menn og kvinner.



Figur 2.1: Predikert sannsynlighet for å ha høy lønn ut fra regresjonsmodellen i Tabell 2.2. Arbeidslivsundersøkelsen 1993.

Vi ser imidlertid også at modellen predikerer negative andeler som har høy lønn blant kvinner som har 0 eller 1 års utdanning utover grunnskole. Dette gir ikke mening. Andelen kvinner som har høy lønn kan ikke blir mindre enn null. En mulighet er å fortolke negative sannsynligheter som 0 og sannsynligheter høyere enn 1 som 1, men dette er også betenkelig dersom antallet enheter som faller utenfor er relativt stort.

Tabell 2.3: Fordeling av predikerte verdier på bakgrunn av regresjonsmodellen i Tabell 2.2. Arbeidslivsundersøkelsen 1993.

	Value	*Frequency*	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	-0,04353	273	8,4	8,7	8,7
	-0,01504	519	16	16,5	25,1
	0,04195	281	8,7	8,9	34
	0,05278	274	8,5	8,7	42,7
	0,08127	356	11	11,3	54
	0,09894	192	5,9	6,1	60,1
	0,13826	618	19,1	19,6	79,7
	0,15593	159	4,9	5	84,8
	0,19524	197	6,1	6,3	91
	0,21291	24	0,7	0,8	91,8
	0,25223	111	3,4	3,5	95,3
	0,30922	137	4,2	4,3	99,7
	0,3947	11	0,3	0,3	100
	Total	3152	97,3	100	
Missing	System	87	2,7		
Total		3239	100		

Hvor stort problemet med meningsløse prediksjoner er, avhenger av fordelingen på den avhengige variabelen og hvilke variabler vi inkluderer i regresjonsmodellen. Generelt er det slik at jo mer skjevfordelt den avhengige er, jo større er problemet. Dersom vi lagrer de predikerte sannsynlighetene fra den lineære sannsynlighetsregresjonen ovenfor, kan vi se hvor stort problemet er i vårt eksempel.

Tabellen ovenfor viser at ganske nøyaktig en fjerdedel av de predikerte sannsynlighetene er lavere enn 0. Det er ingen predikerte sannsynligheter over 1. Problemet med meningsløse prediksjoner representerer med andre et betydelig problem i denne analysen.

2.4.3 Problemet med funksjonell form

Grunnen til at den lineære modellen predikerer negative sannsynligheter er nettopp at den er lineær, dvs. at endringen i sannsynlighet/andel som følge av endringer i uavhengig variabel er lik uansett utgangspunkt. Denne forutsetningen kan i svært mange sammenhenger ikke forsvares hverken teoretisk eller empirisk. Er det f.eks. rimelig å anta at sannsynligheten for å ha høy lønn stiger like mye med økt utdanning uavhengig av hvor stor denne sannsynligheten er på forhånd? I det neste kapitlet skal vi gå nærmere inn på en bestemt funksjonell form som ofte kan forsvares teoretisk og empirisk, nemlig den såkalte S-kurven.

Den funksjonelle formen er på mange måter det viktigste ankepunktet mot lineær sannsynlighetsregresjon, faktisk viktigere enn de mer tekniske svakhetene som er nevnt ovenfor. Grunnen til dette er at man med dette retter oppmerksomheten mot at valg av analysemodell ikke bare er et tekniske spørsmål, men i høyeste grad et teoretisk spørsmål. En velger ikke logistisk regresjon først og fremst på grunn av de tekniske problemene med lineær sannsynlighetsregresjon, men først og fremst fordi den logistiske modellen (og dens fetter probitmodellen) innebærer å konstruere mer realistiske og teoretiske relevante modeller over bestemte sosiale fenomener. Det er viktig å ha i mente at logistisk regresjon (eller probitregresjon) ikke er et godt alternativ i alle situasjoner.

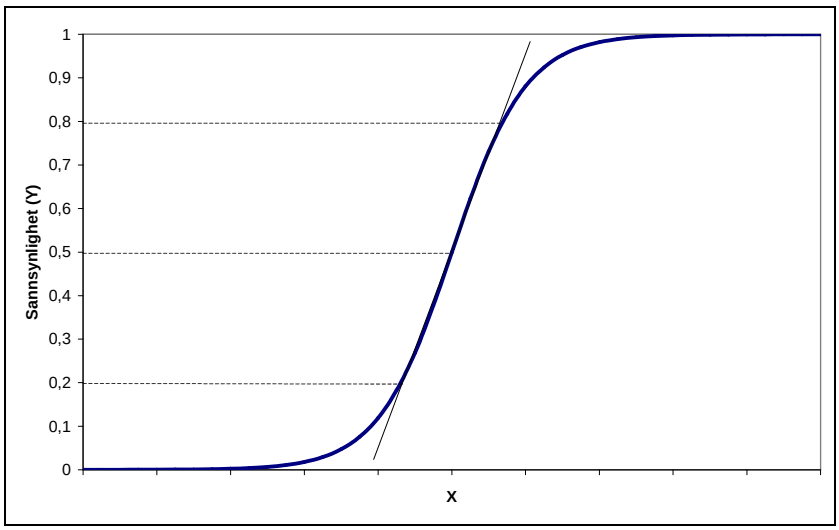
3 Logistisk regresjon med dikotom avhengig variabel

En alternativ måte å foreta analyse av dikotome avhengige variabler er å bruke f.eks. logistisk regresjon eller probitregresjon. Disse to analyseteknikkene har ikke de svakhetene som er innebygd i lineær sannsynlighetsregresjon og tilbyr en funksjonell form som i mange tilfeller er mer realistisk og teoretisk relevant. Prisen man må betale er imidlertid at resultatene ikke lenger gir umiddelbar mening og følgelig kan være vanskelig å fortolke og ikke minst formidle til andre. Framstillingen nedenfor konsentrerer seg utelukkende om logistisk regresjon og tilstreber å gi både en intuitiv og en mer formell innføring i metoden.

3.1 En innledende, intuitiv innføring i prinsippet bak logistisk regresjon

3.1.1 S-kurven og dens anvendelser

I en rekke sammenhenger benyttes S-kurven for å beskrive sosiale fenomener. Denne funksjonelle formen illustrert i figuren nedenfor.



Figur 3.1: Illustrasjon av en S-kurveformet sammenheng mellom sannsynlighet (Y) for et fenomen og en variabel X.

S-kurven er slik at sammenhengen mellom avhengig og uavhengig variabel er svak når den avhengige variabelen i utgangspunktet har lave verdier. Etterhvert som utgangsverdien på den avhengige variabelen øker blir sammenhengen sterkere fram til et visst punkt (midt på kurven). Deretter avtar sammenhengen i styrke igjen.

Endringen i Y som følge av en enhets endring i X er med andre ord minst i begge «halene» på kurven. Det er også verdt å merke seg at kurven aldri når grenseverdiene, selv om det er litt vanskelig å se ut fra figuren. Sannsynligheten (Y) aldri blir null, selv om kurven nærmer seg dette punktet når X går mot en uendelig stor negativ verdi. Det samme gjelder verdien 1 på Y. Når X antar et uendelig stort positivt tall, vil kurven nærme seg 1, men aldri helt nå punktet. Matematisk skrives dette på følgende måte:

$$\lim_{X\to\infty} Y = 1 \quad \text{og} \quad \lim_{X\to-\infty} Y = 0$$

Det første av disse uttrykkene sier er at når variabelen X går mot uendelig positive verdier ($X \to \infty$) så går variabelen Y stadig nærmere en grenseverdi på 1 (lim X = 1), men når aldri denne verdien. Det andre uttrykket sier at når variabelen X går mot uendelig negative verdier ($X \to -\infty$) så går variabelen Y stadig nærmere en grenseverdi på 0 (lim X = 0), men når heller aldri denne verdien.

Figuren ovenfor viser en positivt sammenheng mellom X og Y, nemlig at Y stiger når X stiger. En kan også tenke seg en negativ sammenheng. Da vil kurven gå mot verdien 1 på Y-aksen etterhvert som X antar stadig lavere verdier og mot 0 etterhvert som X øker i verdi. Kurven i figuren blir følgelig speilvendt ved en negativ sammenheng.

Kurvens form kan f.eks. illustreres ved hjelp av sammenhengen mellom tid og læring. Til å begynne med kreves det forholdsvis mye innsats for å tilegne seg kunnskaper på et felt. Effekten av innsatsen er med andre ord forholdsvis liten. Etterhvert som grunnlaget er lagt øker imidlertid læringseffekten drastisk og en får mye igjen for studieinnsatsen. På et senere stadium går imidlertid læringen tregere igjen og det en tilegner seg er mindre i forhold til innsatsen enn tidligere. Det er også slik at selv om en skulle bli en ekspert på området, så vil en aldri sitte med full innsikt. En vil med andre ord aldri helt nå målet (verdien 1).[4]

En rekke sosiale fenomener kan sies å oppføre seg på tilnærmet samme måte: spredning av informasjon som følge av informasjonstiltak eller effekt av sosiale og politiske tiltak som følge av økonomiske ressurser som stilles til rådighet. Formen på kurven kan også knyttes til resonnementer om grensenytte. Mens tilleggsnytten av å øke innsatsen på et område er økende fram til midtpunktet på kurven (der den er brattest), avtar tilleggsnytten gradvis når man øker innsatsen utvoer dette punktet. Et siste eksempel er det vi allerede har benyttet og som vi skal benytte videre, nemlig sammenhengen mellom utdanning og høy lønn. Blant dem som har lav utdanning er andelen som har høy lønn (=sannsynligheten for å ha høy lønn). Mer utdanning øker andelen eller sannsynligheten for å ha høy lønn, men effekten er relativt liten å begynne med, etterhvert blir effekten av mer utdanning større, men effekten blir gradvis mindre etter å ha nådd et vist punkt. Vi ser at det i mange tilfeller er mulig å knytte teoretiske resonnementer til sosiale sammenhenger som stemmer overens med S-kurven ovenfor. Dette er et viktig utgangspunkt for logistisk regresjon.

Et annet viktig poeng i forbindelse med kurven ovenfor bør poengteres. Figuren viser en logistisk sammenheng mellom en tenkt uavhengig variabel (X) og sannsynligheten for at et fenomen inntreffer (Y). Vi ser at Y varierer mellom 0 og 1, slik vi tidligere påpekte at sannsynligheter eller andeler må gjøre.[5] Vi ser også at den logistiske kurven i området mellom 0,2 og 0,8 på Y-aksen, kan beskrives som en rett kurve. Det forteller oss at dersom hovedtyngden av materialet ligger i dette området, dvs. at vi ikke har en svært skjevfordelt variabel, så kan lineær sannsynlighetsregresjon benyttes som en tilnærming. Vi har imidlertid fortsatt problemet med heteroskedastisitet dersom vi ikke benytter WLS som metode (se avsnitt 2.4.1 ovenfor).

[4] Skal vi følge logikken helt ut, antyder kurven også at det ikke er mulig å være komplett idiot på et område (dvs. ha verdien 0), selv om en kan komme farlig nær.

[5] Strengt er formuleringen ovenfor noe upresis. Sannsynligheter må befinne seg i intervallet 0-1, mens andeler som oftest uttrykkes i prosent og følgelig må befinne seg i intervallet 0-100 prosent. Resonnementene ovenfor gjelder imidlertid også for andeler. S-kurven vil da mot 0 på den ene siden og mot 100 på den andre siden. Det er med andre ord kun skalaen på y-aksen som endrer seg når en opererer med andeler.

3.1.2 Hvordan få en avhengig variabel uten grenser

En av grunnene til at lineær sannsynlighetsregresjon ikke faller heldig ut er at denne metoden behandler alle avhengige variabler som kontinuerlige variabler med i prinsippet ubegrenset utfallsrom. Dette gjelder i prinsippet alle former for regresjon. Veien å gå er følgelig å få formulert den avhengige variabelen i regresjonsmodellen slik at den ikke lenger er begrenset til intervallet [0-1]. Klarer vi dette, slipper vi å bekymre oss for utfallsrommet for prediksjonene til modellen.

Figuren ovenfor antyder hvordan dette kan gjøres. Den viser sammenhengen mellom en uavhengig variabel X og en avhengig variabel Y. Legg merke til at mens Y-variabelen ikke kan ha verdier høyere enn 1 eller lavere enn 0, gjelder det ingen tilsvarende begrensninger for X-variabelen. Den kan i prinsippet anta alle reelle verdier, fra pluss uendelig til minus uendelig. Sammenhengen som S-kurven i figuren illustrerer kan med andre ord brukes til å omforme en begrenset variabel (Y) til en ubegrenset variabel (X). Vi kan tenke oss at vi transformerer en sannsynlighetsvariabel (Y) som varierer mellom 0 og 1 til en variabel (X) som ikke har noen begrensninger når det gjelder hvilke verdier den kan anta. En slik variabel kan uten problemer benyttes som avhengig variabel i regresjon.

3.2 En mer formell utledning av den logistiske kurven – odds og log-odds

Ovenfor har jeg presentert det en kunne kalle for det «idemessige» grunnlaget for den logistiske kurven og logistisk regresjon. I dette underkapitlet gjennomgår jeg hvordan vi matematisk kan utlede den logistiske regresjonsmodellen. Utgangspunktet er at vi har en dikotom avhengig variabel Y som varierer mellom 0 og 1. I tråd med hva vi kom fram til ovenfor må denne transformeres til en variabel som i prinsippet ikke har noen begrensninger knyttet til hvilke verdier den kan anta. Dette skjer i to trinn. Det første trinnet er å fjerne grensen for hvor store positive verdier den avhengige variabelen kan anta ved omforme variabelen til odds. Det andre trinnet er å fjerne grensen for hvor store negative verdier variabelen ved å ta logaritmen av oddsen. Nedenfor gjennomgås hvert av disse trinnene i detalj.

3.2.1 Fjerne øvre grense - odds

Det første vi må gjøre er å fjerne den øvre grensen på den avhengige variabelen slik at den i prinsippet kan anta uendelige positive verdier. Det gjøres ved å omforme den avhengige variabelen til et forholdstall kalt *odds*. Oddsen uttrykker et blandingsforhold, dvs. forholdet mellom sannsynligheten for at noe inntreffer (p) mot sannsynligheten for at det ikke inntreffer (1-p). Formelen er: [6]

$$\text{Odds} = \frac{p}{1-p}$$

Eksempel:
Dersom sannsynligheten for at menn har kjøpt svart arbeid er 0,4, er sannsynligheten for at de ikke har kjøpt svart arbeid 1-0,4 = 0,6. Oddsen uttrykker forholdet mellom disse to sannsynlighetene.

[6] Oddsen kan naturligvis også beregnes på bakgrunn av prosentandeler. Siden prosentandeler varierer mellom 0 og 100 blir formelen for odds da (p uttrykker prosentandel):

$$\text{Odds} = \frac{p}{100-p}$$

$$\text{Odds} = \frac{p}{1-p} = \frac{0,4}{1-0,4} = \frac{0,4}{0,6} = 0,67$$

Oddsen 0,67 (2/3) sier at det for hver andre mann som har kjøpt svarte tjenester er det tre menn som ikke har gjort det. Jo høyere oddsen er, jo flere er det som har kjøpt svart i forhold til de som ikke har kjøpt svart. En odds på 2 sier at det for hver andre mann som har kjøpt svart er en som ikke har gjort det.

Odds
Oddsen sier noe om blandingsforholdet blant enhetene med hensyn til en variabel. Andelen eller antallet enheter som har verdien 1 på den avhengige variabelen ses i forhold til andelen eller antallet enheter som har verdien 0. Dette er ikke intuitivt like lett å forstå. La oss derfor illustrere det med to relativt hverdagslige eksempler.

Det første eksempelet er kanskje det minst hverdagslige for de fleste av oss. Odds brukes ofte som et mål på vinnersjanser, f.eks. ved hesteveddeløp. Oddsen forteller hvor mye en vil vinne for hver krone en satser, dersom den hesten en satser på vinner. En hest med høye odds er følgelig en hest som har små vinnersjanser, men som ville gi deg en god sum penger dersom den kommer først i mål. Dersom oddsen er 100:1 («hundre til en»), innebærer det at en får hundre kroner for hver krone en har satset dersom hesten vinner.

Det andre eksempelet er nok mer hverdagslig, nemlig blanding av saft. På kjøpesaft er det ofte oppgitt hvor mye saft en skal blande i forhold til vann. Dersom en skal blande en del saft og fire deler vann, innebærer det at blandingsforholdet er 1:4 («en til fire») med saft. Jo høyere blandingsforholdet er, jo mer saft er det i forhold til vann og jo sterkere er saften.

Tilsvarende er det med eksempelet ovenfor. Jo høyere oddsen er for menn når det gjelder å ha kjøpt svarte tjenester, jo flere er det som har kjøpt svart i forhold til dem som ikke har gjort det.

Generelt er det slik at dersom det er like sannsynlig at noe inntreffer som at det ikke inntreffer, er oddsen for at det inntreffer lik 1. Er oddsen større enn 1, er det mer sannsynlig at fenomenet inntreffer enn at det ikke gjør det. Er oddsen mindre enn 1, innebærer det at det er mer sannsynlig at fenomenet ikke inntreffer enn at det inntreffer. Vi kan også illustrere oddsbegrepet med andeler. En odds på 1 sier at andelen som (eller antallet) som har en egenskap e.l. er like stor som andelen som ikke har denne egenskapen. Er oddsen større enn 1, innebærer det at det er en større andel enheter som har egenskapen i forhold til andelen som ikke har den. Er oddsen mindre enn 1 er andelen som har egenskapen mindre enn andelen av enhetene som ikke har den.

Vi ser også at oddsen kan bli uendelig stor etterhvert som p (sannsynligheten/andelen) blir større:

p = 0,9	→	odds = 0,9/0,1 = 9,
p = 0,99	→	odds = 0,99/0,01 = 99,
p = 0,999	→	odds = 0,999/0,001 = 999 etc.

Det innebærer at vi ved å omforme den avhengige variabelen fra rene sannsynligheter (slik som i sannsynlighetsregresjon) til odds har oppnådd å fjerne den øvre grensen for variabelen. Oddsen kan bli uendelig stor, men den kan ikke bli mindre enn null. Det gjenstår altså å fjerne den nedre grensen.

3.2.2 Fjerne nedre grense – log odds

Det neste skrittet er derfor å fjerne den nedre grensen for variabelen. Vi fjerner den nedre grensen ved å ta den naturlige logaritmen av oddsen. Oddsen varierer mellom 0 og positivt uendelig, slik at vi kan alltid ta logaritmen av denne. Å ta logaritmen av oddsen gir et interessant resultat:

- Når oddsen er lik 1, dvs. at det er like sannsynlig at egenskapen er tilstede som at den ikke er det, så er logaritmen lik 0.
- Er oddsen større enn 1, dvs. at det er mer sannsynlig at egenskapen er tilstede enn at den ikke er det, er logaritmen positiv.
- Er oddsen mindre enn 1, dvs. at det er mer sannsynlig at egenskapen ikke er tilstede enn at den er tilstede, så er logaritmen negativ.

Det uttrykket vi får når vi tar logaritmen av oddsen, kalles logiten, som betegnes med L:

$$L = \ln\left(\frac{p}{1-p}\right)$$

Eksempel:
I vårt eksempel var sannsynligheten for å ha kjøpt svarte tjenester 0,2 for kvinner og 0,4 for menn. Vi kan følgelig beregne log odds eller logiten for de to gruppene:

For kvinner: $$L = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{0,2}{1-0,2}\right) = \ln(0,25) = -1,386$$

For menn: $$L = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{0,4}{1-0,4}\right) = \ln(0,67) = -0,405$$

Disse logitene har ingen enkel substansiell fortolkning. Vi kan imidlertid legge merke til at rangeringen mellom menn og kvinner fortsatt er den samme etter transformeringen. Kvinner har den laveste sannsynligheten for å ha kjøpt svart arbeid og har også den laveste logiten. Transformeringen av sannsynligheter endrer ikke rekkefølgen mellom observasjonene.

Naturlig logaritme
For den som ikke er matematisk bevandret er det umiddelbart ikke lett å forstå hva den naturlige logaritmen er. Formelt er den naturlige logaritmen (ln) til et tall det tallet som en må opphøye det naturlige tallet *e* (=2,718...) i for å få utgangstallet. Eks: Den naturlige logaritmen til 10 er 2,3026... fordi en må opphøye *e* i 2,3026...'nte potens for å få tallet 10.

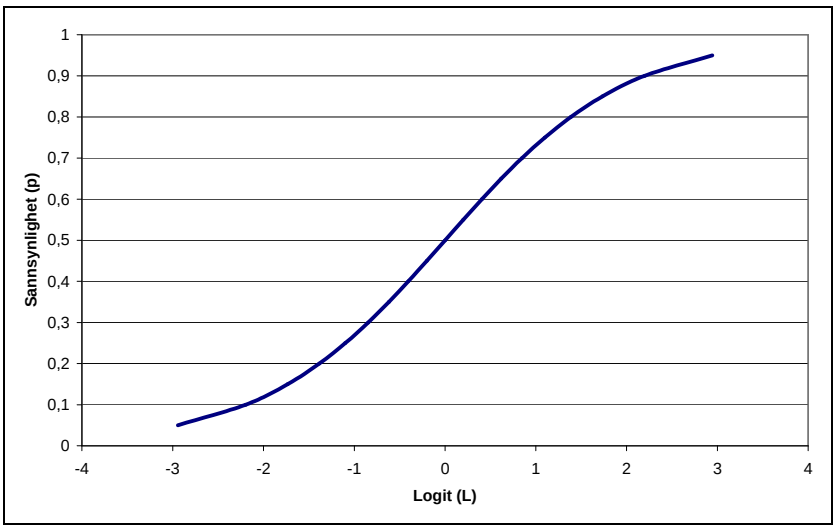
I stedet for å fortape oss i matematiske detaljer, er det en del egenskaper ved den naturlige logaritmen som er viktige i vår sammenheng. For det første medfører det å ta logaritmen av en rekke med tall at *rangeringen mellom tallene opprettholdes*. Det laveste tallet før transformeringen er fortsatt det laveste tallet etterpå, osv. For det andre er *logaritmen til tallet 1 lik 0*. Tall mellom 1 og 0 gir negativ logaritme, mens tall som er høyere enn 1 gir positiv logaritme. Det er ikke mulig å ta logaritmer av tallet null eller negative tall.

Uttrykket i formelen for logit eller log odds har ingen nedre eller øvre grense. Når p (sannsynligheten for at enomenet inntreffer) nærmer seg 0, går L mot minus uendelig:

p = 0,1	→	L=ln(0,1/(1-0,1))	= - 2,20
p = 0,01	→	L=ln(0,01/(1-0,01))	= - 4,60
p = 0,001	→	L=ln(0,001/(1-0,001))	= - 6,91

Ved å transformere sannsynlighetene til log odds har vi med andre ord fjernet både den øvre og nedre grensen for den avhengige variabelen. Logiten kan anta både uendelig store negative verdier og uendelig store positive verdier.

Dersom vi plotter sammenhengen mellom logiten og den underliggende sannsynligheten som den er beregnet på grunnlag av, får vi følgende sammenheng:



Figur 3.2: Illustrasjon av den logistiske sammenhengen mellom sannsynligheten (p) for et fenomen og logiten (L).

Dette er jo faktisk den S-kurven vi gjorde rede for tidligere. Mens sannsynligheten (p) beveger seg mellom 0 og 1, beveger logiten seg fra minus uendelig til pluss uendelig. Det er logiten vi bruker som avhengig variabel i logistisk regreasjon fordi denne ikke er begrenset til å bevege seg innen et bestemt intervall.

Den modellen vi benytter i logistisk regresjon kan formuleres i likningen nedenfor. I likningen betegner *L* logiten, dvs. logaritmen av oddsen; b_0 er koeffisienten for konstanten, dvs. logiten når alle de uavhengige variablene i modellen har verdien 0; x_1-x_n er de uavhengige variablene; b_1-b_n er koeffisientene for de uavhengige variablene, mens *e* står for feilleddet.

$$L = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + \ldots + b_n x_n + e$$

Likningen viser at logaritmen av oddsen (logiten) er en lineær funksjon av et sett uavhengige variabler. Etter å ha beregnet denne modellen, vil vi ikke få som resultat koeffisienter som forteller oss end-

ringen i sannsynlighet som følge av endringer i de uavhengige variablene, men derimot koeffisienter som gir endringen i logiten som følge av endringer i de uavhengige variablene.

3.3 Framgangsmåte og resultater ved logistisk regresjon

3.3.1 Fra en dikotom avhengig variabel til logiter

Jeg har vist hvordan sannsynligheter kan transformeres til en logitvariabel uten nedre og øvre grense. Denne løsningen innebærer imidlertid et tilleggsproblem: Vi kjenner normalt ikke for enhetene i undersøkelsen, dvs. logaritmen av forholdet mellom sannsynligheten for å ha verdien 1 og sannsynligheten for å ha verdien 0 på den avhengige variabelen. Logiten er ukjent for oss fordi vi ikke har oppgitt de underliggende sannsynlighetene. Det eneste observasjonen vi har er en dikotom variabel med verdiene 0 og 1. Vi har data på individnivå, mens en for å beregne sannsynligheter, odds og logit må gruppere enhetene i grupper for å beregne andeler.

En mulighet er faktisk å gruppere enhetene i grupper etter hvilke verdier de har på de uavhengige variablene, deretter estimere andelene i disse gruppene som har verdien 1 på den avhengige variabelen, videre beregne log-odds for så å gjennomføre en regresjonsanalyse på de grupperte dataene. Denne framgangsmåten har imidlertid sine begrensninger i og med at vi er avhengig av at det er et rimelig antall enheter innen hver kategori. Jo flere uavhengige variabler vi trekker inn og jo flere verdier disse variablene har, jo vanskeligere er det å tilfredsstille denne forutsetningen.

Eksempel:
La oss likevel se litt nærmere på denne framgangsmåten. Vi går tilbake til eksempelet i avsnitt 2.4.2 ovenfor med lønnsnivå som dikotom avhengig variabel. Hvis vi forenkler eksemplet og kun trekker inn utdanning som uavhengig variabel, kan vi beregne sannsynligheter, odds og logit for å ha høy lønn innenfor de ulike utdanningskategoriene slik det er gjort i tabellen nedenfor.

Tabell 3.1: Sannsynligheter, odds og logit for å ha høy lønn avhengig av utdanningslengde. Arbeidslivsundersøkelsen 1993.

	Lønn					
Utdanning i år etter grunnskole	lav (y=0)	høy (y=1)	Sum	sannynlighet for å ha høy lønn (p)	odds (p/(1-p))	logit (ln(p))
0	532	15	547	0,027	0,028	-3,569
1	847	28	875	0,032	0,033	-3,409
3	817	82	899	0,091	0,100	-2,299
5	339	50	389	0,129	0,147	-1,914
7	239	31	270	0,115	0,130	-2,042
9	89	72	161	0,447	0,809	-0,212
12	6	5	11	0,455	0,833	-0,182

For å finne sannsynlighetene for å ha høy lønn deles antall enheter med høy lønn på summen av enheter i gruppen. For enheter med 0 års utdanning utover grunnskole blir dette 15/547 = 0,027. Omformingen til odds og logiter skjer i henhold til de formler som er gjennomgått ovenfor. Logitene kan nå brukes som avhengig variabel i en ordinær regresjon med utdanning som uavhengig. Det gir følgende resultat:

Tabell 3.2: Beregning av logiter for å ha høy lønn med. Ordinær lineær regresjon. Arbeidslivsundersøkelsen 1993.

	Koeffisienter	*Standardfeil*	*t-Stat*	*P-verdi*
Utdanning i år	0,298	0,042	7,172	0,001
Konstant	-3,520	0,276	-12,766	0,000

Dette resultatet kan sammenliknes med resultatet fra en logistisk regresjonsanalyse:

Tabell 3.3: Beregning av logiter for å ha høy lønn. Logistisk regresjon. Arbeidslivsundersøkelsen 1993.

	Koeffisienter	*Standardfeil*	*z-Stat*	*P-verdi*
Utdanning i år	0,317	0,022	14,246	0,000
Konstant	-3,514	0,123	-28,581	0,000

Vi ser at koeffisientene for de modellene i dette tilfellet er svært like. Koeffisientene i de to modellene kan også fortolkes på samme måte (vi kommer tilbake til fortolkningen nedenfor). Standardfeilene og t-testene avviker imidlertid temmelig mye fra hverandre i de to modellene. Beregningen av disse i den første modellen er imidlertid basert på 7 observasjoner fordi vi har gruppert enhetene avhengig av hvilken verdi de har på den avhengige variabelen.

Hensikten med dette eksempelet er ikke å vise at ordinær regresjon med manuelt transformert variabel er et alternativ til logistisk regresjon, men å prøve å illustrere hva som egentlig skjer ved logistisk regresjon.

3.3.2 Beregning av koeffisienter i logistisk regresjon – maximum likelihood

Beregningsmåten ved logistisk regresjon følger samme prinsipp som løsningen for lineær sannsynlighetsregresjon. Som vi så i avsnitt 2.3 får en ved ordinær regresjon (OLS)[7] koeffisienter som kan fortolkes som sannsynligheter fordi modellen beregner gjennomsnittsverdier. En dikotom avhengig variabel med verdiene 0 og 1 ville derfor gi resultater som uttrykker sannsynligheter. Problemet er imidlertid at det ikke er teknisk mulig å sette opp regresjonslikningen på en slik måte at vi kan bruke OLS til å beregne logaritmer av odds. Beregningen av koeffisientene i logistisk regresjon, må derfor skje på en annen måte, ved av såkalt Maximum Likehood Estimation (MLE), eller på dårlig norsk: Estimering av maksimal sannsynlighet.

Ved estimeringen av koeffisienter i ordinær regresjon benyttes minste kvadraters metode. Det innebærer at en estimerer de parametrene som ut fra den spesifiserte modellen gir minst mulig total avstand (Error Sum of Squares) mellom de predikerte verdiene og de observerte verdiene. Vi ønsker totalt sett minst mulig feilprediksjoner.

Maximum Likelihood Estimation arbeider på en noe annen måte. Gitt den modellen vi har spesifisert (se ovenfor), beregnes de estimatene som gjør det mest sannsynlig at vi har fått de observerte y-verdiene (0 og 1). Det formuleres en Maximum Likelihood funksjon som utgangspunkt for dette estimeringsproblemet. Selve beregningen går ut på å finne de estimatene som maksimerer denne funksjonen. Som vi ser av S-kurven er den logistiske regresjonsmodellen ikke-lineær og matematikken er betrakte-

[7] OLS og ordinær regresjon blir brukt som synonymer for regresjon basert på minste kvadraters metode.

lig mer komplisert enn ved ordinær regresjon. Modellen kan ikke løses ved hjelp av algebra og løsningen må derfor finnes numerisk, dvs. en må ved hjelp av algoritmer «prøve» seg fram til løsningen av maksimeringsproblemet. Enkelte statistikkprogrammer, som f.eks. Stata, viser resultatet av hver enkelt iterasjon på utskriften.

Selv om siktemålet for de to estimeringsmetodene er noe forskjellig, er framgangsmåtene egentlig ikke så ulike. Både minste kvadraters metode (OLS) og Mamimum Likelihood Estimation dreier seg om å tilpasse en likning slik at den passer best mulig med det datamateriale vi ønsker å beskrive. Det er også slik at OLS-estimering kan betraktes som et spesialtilfelle av ML-estimering (Gujarati 1988:96).

3.3.3 Regresjonsresultater

På samme måte som lineær regresjon, gir logistisk regresjon estimater koeffisienter for konstantleddet og hver av de inkluderte uavhengige variabler. Siden regresjonsmodellen har følgende form,

$$L = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + ... + b_n x_n + e$$

er fortolkningen av koeffisientene ikke like innlysende som ved ordinær regresjon. Den avhengige variabelen er ikke sannsynligheten for å ha verdien 1 på avhengig variabel som ved sannsynlighetsregresjon, men logaritmen av oddsen for å ha verdien 1, logiten.

Konstanten (b_0) viser gjennomsnittlig logit når alle de uavhengige variablene i modellen har verdien 0. De andre koeffisientene (b_1-b_n) viser hvor mye logiten eller log-oddsen endres når en uavhengig variabel øker med en enhet i verdi og de andre uavhengige variablene holdes konstant.

Eksempel:
Vi foretar den samme analysen som i eksempelet med lineær sannsynlighetsregresjon, men benytter nå logistisk regresjon. Den avhengige variabelen er som før timelønn, dikotomisert i verdiene inntil 140 kr/t (0) og 140 kr/t eller mer (1). De uavhengige variablene er fortsatt kjønn og utdanning. Dette gir følgende modell:

$$\ln\left(\frac{P(Y=1)}{1 - P(Y=1)}\right) = b_0 + b_1 \times \text{Kjønn} + b_2 \times \text{Utdanning} + e$$

Resultatet av en logistisk regresjon basert på denne modellen er vist i tabellen nedenfor. Utskriften er hentet fra statistikkprogrammet SPSS ved å velge **Regression ▸ Logistic...** under **Statistics** på menylinjen. I Stata effektueres logistisk regresjon med kommandoen **.logit**.

Tabell 3.4: Logistisk regresjon med lønnsnivå (dikotomisert) som avhengig variabel og kjønn og utdanning som uavhengige. Arbeidslivsundersøkelsen 1993.

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Kjønn	-1,6026	0,1763	82,6356	1	0,0000	-0,2058	0,2014
Utdanning	0,2987	0,0228	171,5556	1	0,0000	0,2984	1,3482
Konstant	-2,9629	0,1279	536,616	1	0,0000		

Den første kolonnen (B) viser paramterestimatene. Konstanten viser at når kjønn og utdanning har verdien 0, dvs. for menn uten utdanning utover grunnskole, er den estimerte logiten eller logaritmen av oddsen (log odds) for å ha høy lønn –2,96. Kjønnskoeffisienten viser at log odds for å ha høy lønn er i gjennomsnitt –1,60 lavere for kvinner enn for menn, kontrollert for utdanning. Når det gjelder utdanning, er log odds for å ha høy lønn i gjennomsnitt 0,3 høyere for hvert år mer utdanning respondentene i undersøkelsen har. Selv om den rent tekniske fortolkningen av koeffisientene ikke skiller seg fra ordinær regresjon, er det ikke umiddelbart lett forstå hva koeffisientene innebærer. De færreste av oss tenker naturlig i termer av log-odds. I neste underkapittel vil vi derfor se på ulike strategier for å presentere resultater fra logistisk regresjon.

Men først vil jeg raskt beskrive for de andre målene i utskriften ovenfor. I kolonne nummer to (S.E.) finner vi standardfeilene til estimatene og i de tre neste kolonnene (Wald, df og Sig) presenterer resultatene fra signifikanstesting av estimatene basert på standardfeilene. Vi kommer tilbake til disse i underkapittel 4.1.1. I sjette kolonne (R) presenteres et mål på den partielle korrelasjonen mellom den avhengige variabelen og hver av de uavhengige og gir uttrykk for variablenes bidrag til modellen. Dette målet kan sammenliknes med den standardiserte regresjonskoeffisienten i ordinær regresjon. Den siste kolonnen (Exp (B)) angir effekten av de uavhengige variablene i oddsratioer i stedet for log odds. Dette målet kan være nyttig fordi det ofte er lettere å fortolke enn log odds. Jeg kommer tilbake til oddsratioer nedenfor.

3.4 Fortolkninger og presentasjon av koeffisienter i logistisk regresjon

Utskriften ovenfor viste at resultatene fra logstisk regresjon ikke er enkle å fortolke og presentere på en forståelig måte. Her vil jeg diskutere fem mulige løsninger på dette problemet. Den første omgår problemet ved å se bort fra fortolkninger av koeffisientene og retter oppmerksomheten mot fortegnet på koeffisientene og signifikanstester. De fire andre går alle ut på å regne om logiten til enten oddsratioer eller sannsynligheter. I praksis vil en som oftest benytte seg av en kombinasjon av disse tilnærmingene.

3.4.1 Koeffisientens fortegn

Koeffisientene for hver av de uavhengige variablene i vårt eksempel viser hvor mye logaritmen av oddsen for å ha høy lønn, logiten, endres når den uavhengige variabelen øker med en enhet i verdi. Positiv koeffisient innebærer at logaritmen av oddsen for å ha høy lønn øker med høyere verdi på den uavhengige variabelen, negativ koeffisient innebærer at logaritmen av oddsen for å ha høy lønn synker. Siden det å ta logaritmen av en rekke tall endrer forholdet mellom tallene, men ikke rekkefølgen, kan vi fortolke fortegnet på koeffisientene som vi ville ha gjort ved ordinær regresjon:

- Positivt fortegn innebærer positiv sammenheng. Oddsen (og følgelig sannsynligheten/andelen) øker med høyere verdi på den uavhengige variabelen.
- Negativt fortegn innebærer negativ sammenheng. Oddsen (og følgelig sannsynligheten/andelen) synker med høyere verdi på den uavhengige variabelen.

Det betyr at vi f.eks. kan fortolke den negative og signifikante kjønnskoeffisienten i modellen ovenfor som støtte for en hypotese om at andelen som har høy lønn er mindre blant kvinner enn blant menn, kontrollert for utdanningslengde. Dersom vi arbeider med et utvalg av respondenter (og det gjør vi som regel), er det ikke nok å bare se på fortegnet, hvorvidt koeffisienten er statistisk signifikant eller ikke må også trekkes inn. Vi kommer tilbake til dette i avsnitt 4.1.

3.4.2 Oddsratioer

De neste metodene tilstreber alle å få fram størrelsen på sammenhengene i modellen. Den første metoden, å beregne oddsratioer, kan lette forståelsen av koeffisientene noe, men har antakelig likevel begrenset nytteverdi. For å komme beregne oddsratioer må en ta antilogaritmen av de estimerte koeffisientene for de uavhengige variablene. Oddsratioer er et mål på effekt/sammenheng slik at det ikke er mulig å beregne oddsratioer for konstantleddet.

Antilogaritme
Antilogaritmen til et tall er det motsatte av logaritmen til et tall. Mens den naturlige logaritmen (ln) til et tall er det tallet som en må opphøye det naturlige tallet *e* (=2,718...) i for å få utgangstallet, er antilogaritmen av et tall, rett og slett resultatet av å opphøye *e* i dette tallet. Som et eksempel er e^1 lik 2,718... Det innebærer også at følgende sammenheng gjelder:

$$e^{\ln(x)} = x$$

Dette innebærer at dersom en tar antilogen av logaritmen av et tall, får man utgangstallet. Å ta antilogaritmen av log-oddsen vil følgelig gi oddsen som resultat.

Det resultatet en får ved å ta antilogaritmen av logitkoeffisientene viser hvor mange *ganger* oddsen endres når verdien på en uavhengig variabel stiger med en enhet. En annen måte å si det på er hva oddsen (sjansen) for å ha verdien 1 på den avhengige variabelen må ganges med når den uavhengige variabelen stiger i verdi.

Dette tallet kalles for *oddsratio*, dvs. ratioen eller forholdstallet mellom odds. Dersom vi tar oddsen når en uavhengig variabel har en bestemt verdi, $p_1/(1-p_1)$, og deler på den tilsvarende oddsen når den uavhengig variabelen er en verdi lavere, $p_0/(1-p_0)$, får vi oddsratioen:

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

I formelen betegner θ oddsratioen, p_0 og p_1 er sannsynlighetene for å ha verdien 1 på avhengig variabel for de to verdiene på den uavhengige variabelen.

Oddsratioer kan aldri bli negative. En oddsratio som er høyere enn 1, innebærer at oddsen stiger med høyere verdi på uavhengig variabel, mens en oddsratio som er lavere enn 1 innebærer at oddsen synker

med høyere verdi på uavhengig variabel. Er oddsratioen lik 1, er det ingen sammenheng mellom oddsen og den uavhengige variabelen. Det man vinner ved omregningen til oddsratioer er at vi nå får et tall for forholdet mellom odds og at vi slipper å tenke i logaritmer. Tar man dette i betraktning er oddsratioer antakelig enklere å fortolke enn log-odds (logit). Men det ville nok være en overdrivelse å hevde at oddsratioer er intuitivt lett forståelig.

Eksempel:
Tar vi f.eks. kjønnskoeffisienten i regresjonsutskriften ovenfor, finner vi antilogartimen av logiten ved å opphøye *e* i estimatet (-1,6026). $e^{-1,6026}$ gir 0,2014. Det innebærer at forholdet mellom oddsen for å ha høy lønn for kvinner (kjønn=1) og den tilsvarende odds for menn (kjonn=0) er 0,2. Oddsratioen for kjønnskoeffisienten er mindre enn 1. Det innebærer at oddsen for å ha høy lønn er lavere blant kvinner enn blant menn. Sammenhengen er med andre ord negativ. Dette kan fortolkes som at kvinnelige yrkestakeres odds (sjanse) for å ha høy lønn kun utgjør en femtedel av mannlige yrkesdeltakeres sjanse for å ha høy lønn.

Den andre koeffisienten i eksempelet ovenfor gjelder utdanningseffekten. Hvis vi regner antilogen av koeffisienten får vi 1,3481. Oddsratioen er større enn 1, noe som viser at sammenhengen er positiv. Det vil si at for hvert år mer med utdanning, øker oddsen for å ha høy lønn med 1,35 ganger.

Dersom en ganger oddsration med 100 får en endringen i odds uttrykt i prosent. Rettere sagt, en får da et uttrykk for hvor mye den nye oddsen utgjør i prosent av den opprinnelige når den uavhengige variabelen øker med en enhet. Anvendt på eksempelet ovenfor kan vi tolke kjønnskoeffisienten som at oddsen for å ha høy lønn blant kvinner er 20 % av den tilsvarende oddsen for menn, kontrollert for utdanning. Når det gjelden utdanning er oddsen for å ha høy lønn 35 % høyere, for hvert år mer med utdanning en har utover grunnskole, kontrollert for kjønn.

Oddsratioer kan, som vi har sett, enkelt beregnes ved å ta antilogen av koeffisientene ved logistisk regresjon. Legg imidlertid merke til at utregningen egentlig er unødvendig når en bruker SPSS fordi oddsratioen oppgis i den siste kolonnen i utskriften. Dersom en bruker statistikkpakken Stata kan en få resultatet oppgitt i oddsratioer ved å benytte kommandoen **.logistic** i stedet for **.logit**.

3.4.3 Sannsynligheter/andeler

Selv om oddsratioer er lettere å forstå enn log-odds, er det nok et stort behov for mer lettilgjengelig presentasjon av resultater fra logistisk regresjon. Enda enklere enn å forstå enn oddsratioer er naturligvis sannsynligheter. Men i og med at den logistiske modellen er ikke-lineær, kan ikke koeffisientene uten videre gjøres om til sannsynligheter. Effekten av en variabel omregnet til sannsynlighet vil nemlig avhenge av hvor stor utgangssannsynligheten er. Det ses lett av S-kurven som ble presentert tidligere. Vi ser at effekten er størst omtrent midt på kurven mens den avtar i begge halene.

En mulighet er imidlertid å regne ut sannsynligheter eller andeler for gitte kombinasjoner av verdier på de uavhengige variablene. Å regne ut sannsynligheter ut fra log odds innebærer å gå motsatt vei i forhold til da vi gjorde sannsynligheter om til log odds. Formelen for å regne ut sannsynligheter kan utledes ved å omforme litt på likningen for en logistisk regresjonsmodell:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + ... + b_nx_n$$

$$\Updownarrow$$

$$\left(\frac{p}{1-p}\right) = e^{(b_0+b_1x_1+...+b_nx_n)}$$

$$\Updownarrow$$

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+...+b_nx_n+e)}}$$

Vi finner sannsynligheten for å ha verdien 1 på den avhengige variabelen for en gitt kombinasjon av verdier på de uavhengige variablene ved å dele 1 på summen av 1 og antilogaritmen av minus beregnet log odds. For å klare opp i denne noe tungvinte beskrivelsen kan vi ta for oss hvert enkelt element. Log oddsen er rett og slett det beregnede resultatet når en setter inn konkrete verdier for de uavhengige variablene i den likningen som er resultatet av å effektuere regresjonsmodellen. Etter å beregnet log oddsen endres fortegnet på resultatet, slik at en positiv log odds blir negativ og omvendt. Til slutt tas antilogen av dette resultatet. Da får en det som i formelen ovenfor er benevnt:

$$e^{-(b_0+b_1x_1+...+b_nx_n+e}$$

Resultatet settes inn i formelen ovenfor, og sannsynligheten kan beregnes.

Eksempel:
Vi tar fortsatt utgangspunkt i den logistiske regresjonsmodellen som ble presentert ovenfor. Først estimerer vi andelen som har høy lønn blant *kvinner uten utdannelse utover grunnskole*. Dette gjøres i tre trinn:

(1) Beregner *log odds* for kvinner uten utdannelse utover grunnskole:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + ... + b_nx_n = -2{,}96 + (1 \times -1{,}60) + (0 \times 0{,}30) = -4{,}56$$

(2) Tar *antilogaritmen* av minus log oddsen:

$$e^{-(b_0+b_1x_1+...+b_nx_n+e)} = e^{4{,}56} = 95{,}5835$$

(3) Beregner andelen som har høy lønn:

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+...+b_nx_n+e)}} = \frac{1}{(1+95{,}5835)} = 0{,}0104$$

Estimert andel som har høy lønn blant kvinner uten utdannelse utover grunnskole er 1 prosent. Alternativt kan vi si at sannsynligheten for å ha høy lønn for kvinner uten utdannelse utover grunnskole er 0,01.

Deretter estimerer vi andelen som har høy lønn blant *menn uten utdannelse utover grunnskole*:

(1) Beregner log-odds for menn uten utdannelse utover grunnskole:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + ... + b_n x_n = -2{,}96 + (0 \times -1{,}60) + (0 \times 0{,}30) = \text{-}2{,}96$$

(2) Tar antilogen av minus log-oddsen:

$$e^{-(b_0+b_1x_1+...+b_nx_n+e)} = e^{2{,}96} = 19{,}298$$

(3) Beregner andelen som har høy lønn:

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+...+b_nx_n+e)}} = \frac{1}{(1+19{,}298)} = 0{,}0493$$

Estimert andel som har høy lønn blant menn uten utdannelse utover grunnskole er 5%.

Dersom vi foretar de samme beregningene for kvinner og menn med 5 års utdanning utover grunnskole, finner vi at sannsynligheten er 0,0448 (4 prosent) for kvinner og 0,1885 (19 prosent) for menn. Vi kan beregne at prosentdifferansen mellom menn og kvinner når det gjelder andelen som har høy lønn er høyere for dem som har 5 års utdanning utover grunnskole (19% - 4% = 15 %-poeng) enn for dem som utelukkende har grunnskole (5% - 1% = 1 %-poeng).

Selv om forskjellen i sannsynligheter varierer, er oddsratioen mellom kvinner og menn konstant for alle verdier på utdanningsvariabelen:

$$\theta_{\text{Utdanning}=0} = \frac{\frac{p_{\text{kvinner}}}{1-p_{\text{kvinner}}}}{\frac{p_{\text{menn}}}{1-p_{\text{menn}}}} = \frac{\frac{0{,}0104}{1-0{,}0104}}{\frac{0{,}0493}{1-0{,}0493}} = 0{,}2$$

$$\theta_{\text{Utdanning}=5} = \frac{\frac{p_{\text{kvinner}}}{1-p_{\text{kvinner}}}}{\frac{p_{\text{menn}}}{1-p_{\text{menn}}}} = \frac{\frac{0{,}0448}{1-0{,}0448}}{\frac{0{,}1885}{1-0{,}1885}} = 0{,}2$$

Dette innebærer at oddsen for å ha høy lønn for kvinner er 20% av den tilsvarende oddsen for menn, uavhengig av hvilket utdanningsnivå de befinner seg på. Dette er i tråd med hva vi fant i underkapittel 3.4.2 ovenfor. Logistisk regresjon forutsetter med andre ord at effekten målt i oddsratio (og log odds) er konstant, mens effekten målt i endring av sannsynlighet varierer.

3.4.4 Maksimaleffekt og «gjennomsnittseffekt»

De logistiske regresjonskoeffisientene gir også muligheten for å beregne effekten av en uavhengig variabel avhengig av hvor en befinner seg på den logistiske kurven. Utregningen skjer ved hjelp av formelen nedenfor:

$$b \times p(1-p)$$

Ved å gange regresjonskoeffisienten (b) med produktet av sannsynligheten for at den avhengige variabelen har verdien 1 (p) og sannsynligheten at den ikke har verdien 1 (1-p), får man et estimat på hvor mye sannsynligheten endres som følge av en enhets endring i den uavhengige variabelen i dette punktet.

En presentasjon av logistiske regresjonskoeffisienter som kan være hensiktsmessig i noen tilfeller er å beregne den maksimale effekten målt i sannsynlighet som følge av variabelen. Det må være effekten av variabelen der hvor den logistiske kurven er brattest, dvs. der hvor sannsynligheten for å ha verdiene 0 og 1 på den avhengige variabelen er lik (p=0,5 og 1-p = 0,5). Den maksimale effekten kan i henhold til formelen ovenfor beregnes ved å gange regresjonskoeffisienten med produktet av 0,5 og 0,5, dvs. 0,25.

$$\text{Maksimal effekt} = b \times 0{,}5 \times 0{,}5 = 0{,}25b$$

Eksempel:
Dersom vi bruker formelen ovenfor til vårt gjennomgående eksempel, finner vi at de maksimale effektene av kjønn og utdanning er henholdsvis:

Kjønn:	0,25 x -1,60 = - 0,40 eller 40%
Utdanning:	0,25 x 0,30 = 0,08 eller 8%

Den maksimale forskjellen mellom menn og kvinner når det gjelder sannsynligheten for å ha høy lønn er 0,40. En annen måte å si det på er at den maksimale forskjellen mellom menn og kvinner i andel som har høy lønn er 40 prosentpoeng. Ser vi på utdanningsvariabelen, er den maksimale effekten på sannsynlighet for å ha høy lønn lik 0,08. Et års utdanning ekstra kan maksimalt øke sannsynligheten for å få høy lønn med 0,08, eller maksimalt øke andelen som får høy lønn med 8 prosentpoeng. Disse maksimale effektene forutsetter at sannsynligheten for å få høy lønn i utgangspunktet er 0,5 (50/50 fordeling mellom de som har høy lønn og de som ikke har det).

Det er imidlertid verdt å merke seg at det i vår modell innenfor datamaterialets begrensning ikke predikeres sannsynligheter i nærheten av 0,5, slik at de beregnede maksimale effektene har svært liten substansiell betydning i dette tilfelle. En annen mulighet er derfor å beregne effekten når sannsynligheten er lik marginalfordelingen på den avhengige variabelen i utvalget (Sørensen 1989:68). Dette gir et uttrykk for variabelens effekt på sannsynligheten der «tyngdepunktet» av materialet befinner seg, dvs. der for den «typiske» respondenten/enheten. Denne effekten kan noe misvisende kalles den gjennomsnittlige effekt på sannsynligheten.

Eksempel:
I eksempelet ovenfor tilsier marginalfordelingen at p (andel med høy lønn) er omtrent 9% og 1-p (andel med lav lønn) er 91% (jf. avsnitt 2.4.2). Effekten på sannsynligheten av henholdsvis kjønn og utdanning for den «typiske» respondent er da:

Kjønn:	0,09 x (1-0,09) x –1,60 = 0,0819 x –1,60 = - 0,13 eller 13%
Utdanning:	0,09 x (1-0,09) x 0,30 = 0,0819 x 0,30 = 0,025 eller 2,5%

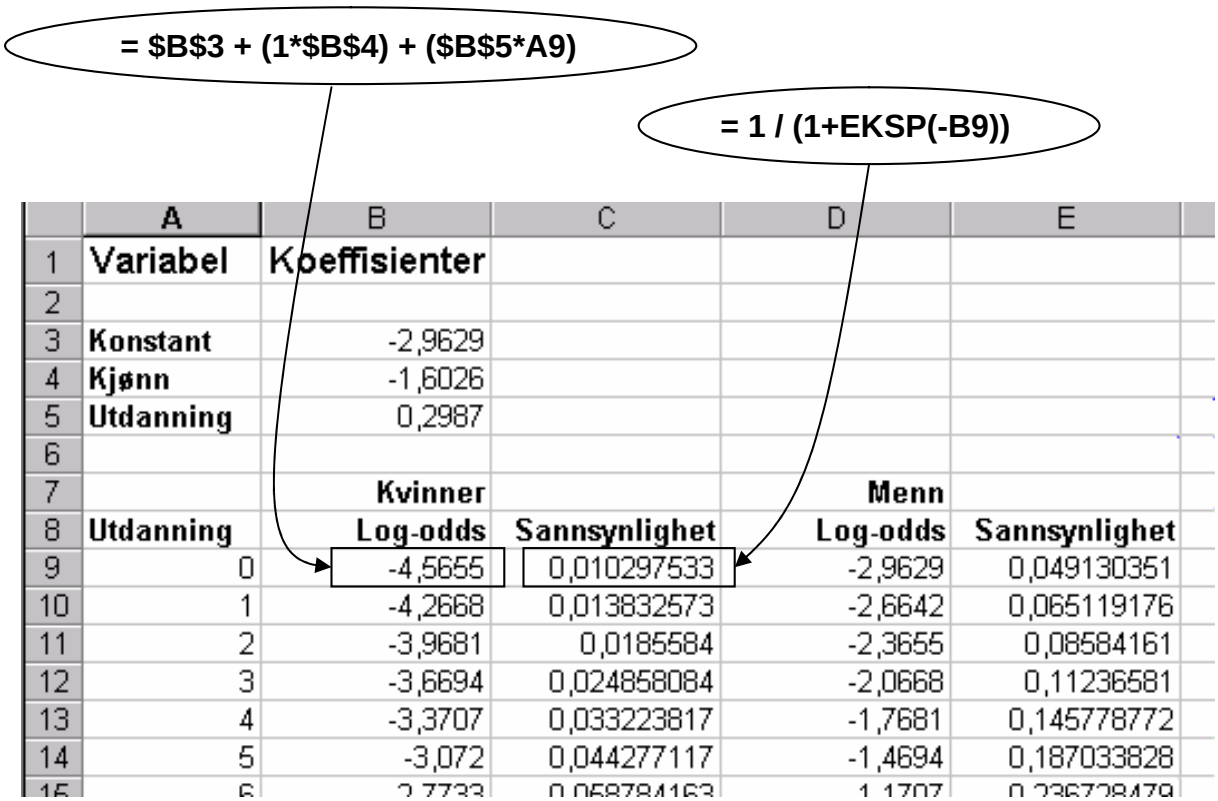
Dette er effektene av de to variablene når verdiene på de uavhengige variablene gir en utgangssannsynlighet som er lik marginalfordelingen (eller gjennomsnittsfordelingen) på den avhengige variabelen. Vi ser at disse effektene er langt lavere enn de respektive maksimaleffektene, fordi variabelen

er svært skjevfordelt og befinner seg langt fra en 50/50 fordeling. Denne «gjennomsnittseffekten» har ofte mer substansiell interesse enn maksimaleffekten.

3.4.5 Grafisk framstilling

Med utgangspunkt i beregnede sannsynligheter kan regresjonsresultatene i avsnitt 3.3.3 presenteres grafisk, på samme måte som vi gjorde i med OLS-resultatene. Dette er kanskje den beste måten for å presentere resultater fra logistisk regresjon fordi den gir et visuelt bilde av sammenhengen som det er vanskelig å forestille seg ut fra logiten, oddsratioer eller beregnede sannsynligheter.

Grafiske presentasjoner kan lages både ved hjelp av grafikkfunksjoner i statistikkprogrammene og ved å bruke elektroniske regneark (f.eks. Excel, Lotus 123). Regneark gir flere muligheter og større fleksibilitet enn de fleste statistikkprogrammer, men krever litt kjennskap til hvordan en lager formler og gjør beregninger. I figuren nedenfor har vi benyttet resultatene fra vårt eksempel til å lage en grafisk framstilling.



	A	B	C	D	E
1	Variabel	Koeffisienter			
2					
3	Konstant	-2,9629			
4	Kjønn	-1,6026			
5	Utdanning	0,2987			
6					
7		Kvinner		Menn	
8	Utdanning	Log-odds	Sannsynlighet	Log-odds	Sannsynlighet
9	0	-4,5655	0,010297533	-2,9629	0,049130351
10	1	-4,2668	0,013832573	-2,6642	0,065119176
11	2	-3,9681	0,0185584	-2,3655	0,08584161
12	3	-3,6694	0,024858084	-2,0668	0,11236581
13	4	-3,3707	0,033223817	-1,7681	0,145778772
14	5	-3,072	0,044277117	-1,4694	0,187033828
15	6	-2,7733	0,058784163	-1,1707	0,236728479

Figur 3.3: Illustrasjon av framgangsmåte for beregning av sannsynligheter i et regneark.

Utgangspunktet for regnearket er at vi ønsker å framstille sammenhengen mellom utdanning og sannsynligheten for å ha høy lønn, kontrollert for kjønn. Vi kan si at regnearket i figuren består av 4 hoveddeler:

(1) variabler og koeffisienter (området A1:B5)
(2) verdier på den uavhengige variabelen som skal varieres (kolonne A fra celle 8 og nedover)
(3) beregning av log odds (kolonne B og D fra celle 8 og nedover)
(4) beregning av sannsynlighet (kolonne C og E fra celle 8 og nedover)

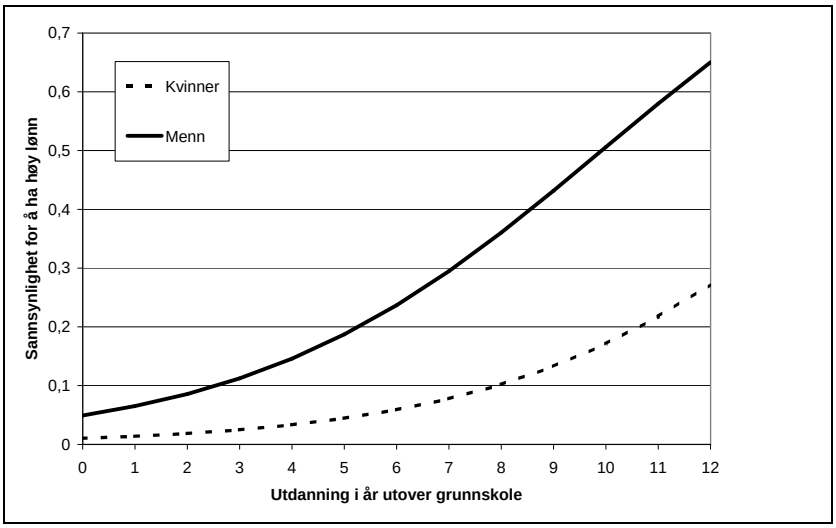
I tillegg til dette kommer selve grafen.

I den første delen legges koeffisientene inn for alle variablene og konstanten slik de kommer fram på utskriften fra f.eks. SPSS eller STATA. I den andre delen angir vi stigende verdier på den uavhengige variabelen som vi ønsker å illustrere effekten av i figuren, nemlig utdanning. Vi må beregne sannsynligheten for å ha høy lønn for ulike verdier på utdanning for å kunne plotte sammenhengen mellom utdanning og lønn.

De to første delene angir input for å estimere sannsynligheter. Selve beregningene er splittet opp i to trinn: først beregnes log odds, deretter beregnes sannsynligheter. Log odds beregnes ved å lage en likning som inkluderer konstant, variabelverdier og koeffisienter. I eksempelet ovenfor er log oddsen summen av konstant (-2,9629), verdien for kjønn (0=mann, 1=kvinne) ganget med kjønnskoeffisienten (-1,6026) og verdien for utdanning (0 og oppover) ganget med utdanningskoeffisienten (0,2987). Vi ser at formlene i Excel består av cellereferanser til de celler hvor de aktuelle tallene befinner seg. Log-odds beregnes for alle verdier på utdanningsvariabelen. I regnearket har vi beregnet sannsynligheter både for menn (kolonne E) og kvinner (kolonne C).

Deretter beregnes sannsynligheten med utgangspunkt i odds. Sannsynligheten framkommer ved å dele 1 på summen av 1 og antilogaritmen av den beregnede log-oddsen (med omvendt fortegn). Ved å lage et regneark som dette kan en automatisere beregningene av sannsynligheter og samtidig være noe sikrere på at beregningene er korrekte.

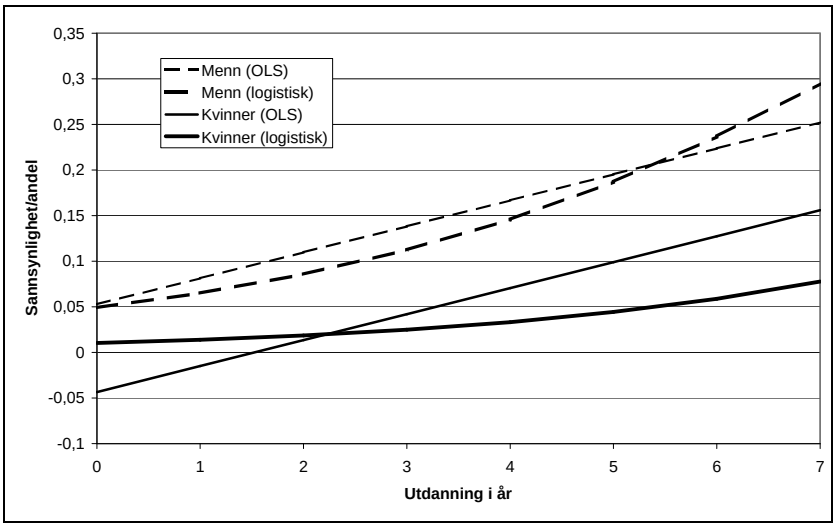
Tallene i kolonne A, C og E kan nå benyttes for å lage en grafisk framstilling av sammenhengen mellom utdanning og lønn. Siden vi har beregnet sannsynligheter for menn og kvinner isolert, får vi to kurver over sammenhengen mellom utdanning og lønn, en for menn og en for kvinner. Vi får dermed fram at kurven stiger brattere for menn enn for kvinner.



Figur 3.4: Sannsynlighet for å ha høy lønn basert på den logistisk regresjonsmodellen i Tabell 3.4. Arbeidslivsundersøkelsen 1993.

Figuren ovenfor viser predikert sannsynlighet for å ha høy timelønn (kr. 140,- eller høyere), avhengig av kjønn og utdanning. Vi ser at den logistiske modellen ikke predikerer negative sannsynligheter, slik OLS gjorde. I tillegg ser vi at kurvene ikke er helt parallelle, slik OLS ga som resultat.

Dersom vi sammenlikner regresjonslinjene fra både OLS og logistisk regresjon, får vi følgende figur:



Figur 3.5: Sammenlikning av predikerte sannsynligheter for å ha høy lønn med henholdsvis lineær sannsynlighetsregresjon og logistisk regresjon. Arbeidslivsundersøkelsen 1993.

Figuren viser at det er relativt stor forskjell mellom de to analyseresultatene, både for menn og kvinner. Forskjellen er likevel størst for kvinner. Grunnen til at kurvene er så forskjellige er at den avhengige variabelen er skjevfordelt. De estimerte andelene er små og nærmer seg null for dem som ikke har utdanning utover grunnskole. Vi befinner oss altså ved begynnelsen av den logistiske kurven hvor effekten målt i sannsynligheter er relativt liten. OLS-kurven bygger imidlertid på forutsetningen om at effekten er like uansett utgangspunkt og to kurvene for følgelig nokså ulik form. Hadde den avhengige variabelen ikke vært så skjevfordelt ville en derimot ha opplevd at kurvene var relativt like.

I en så enkel modell som ovenfor byr det på små problemer å utarbeide grafiske presentasjoner. En kan imidlertid ikke vise sammenhengen for alle variabler dersom modellen er stor og inneholder variabler med mange verdier. Da må en illustrere sammenhengen for utvalgte variabler som er sentrale for problemstillingen. I tillegg må en velge hvilke verdier de andre variablene i modellen skal. Dette er nødvendig for å kunne beregne sannsynlighetene. Siden formen på kurvene er avhengig av utgangspunktet, er det tvilsomt om det å sette de andre variablene til null er noen god løsning, dersom dette ikke er av spesielt teoretisk interesse.

Ofte velges gjennomsnittet eller andre mål sentraltendens på de andre variablene som beregningspunkt. Dette er fordi gjennomsnitt, media eller modus angir steder hvor tyngdepunktet av enhetene i undersøkelsen befinner seg. En annen type kriterium er å velge teoretisk interessante kombinasjoner av verdier på de andre variablene.

Beregninger av sannsynligheter og grafiske framstillinger kan man gjøre i regneark som f.eks. Excel, men en kan også lagre predikerte sannsynligheter i statistikkprogrammer som SPSS og Stata og deretter framstille grafisk sammenhengen mellom de predikerte sannsynlighetene og en uavhengig variabel. Dette er en framgangsmåte som fungerer best ved relativt enkle modeller eller dersom en har datasett med svært mange observasjoner.

4 Signifikanstester og andre statistiske mål ved logistisk regresjon

Hypoteseprøving og mål på modellens tilpasning til dataene er, riktig brukt, viktige verktøy i forskningsprosessen. I prinsippet følger man samme tankegang og de samme framgangsmåter som ved ordinær regresjonsanalyse, men samtidig byr logistisk regresjon på andre utfordringer og muligheter ved at man må ta utgangspunkt i den loglikelihoodfunksjonen som brukes for å estimere regresjonskoeffisientene.

4.1 Signifikanstester

Som oftest er man ikke utelukkende interessert i koeffisientenes størrelse. Ved utvalgsundersøkelser ønsker en også, og kanskje først og fremst, å teste hypoteser om hvorvidt det er sammenhenger mellom en avhengig variabel og et sett av uavhengige variabler. Vi har allerede sett at koeffisientene i logistisk regresjon ikke har en umiddelbart lett tilgjengelig fortolkning. Gjelder dette også for signifikanstester?

4.1.1 Z-test eller Wald-test

For å teste om et estimat er signifikant eller ikke, må det også være mulig å beregne et mål på usikkerhet knyttet til estimatet, ofte kalt standardfeil.[8] Dersom dette er mulig, kan også signifikanstesting gjennomføres. Logistisk regresjonsanalyse gir mulighet for å beregne såkalte *asymptotiske standardfeil (ASE)* som kan benyttes ved statistisk testing av hypoteser om sammenhenger. Hvorvidt en koeffisient, ß, er signifikant eller ikke kan testes som en vanlig t-test eller z-test i ordinær regresjon:

$$z = \frac{\hat{\beta}}{ASE}$$

Denne testoperatoren er tilnærmet normalfordelt når utvalget er stort og ß = 0. Dersom z overskrider en kritisk verdi, f.eks. 1,96 ved 5% signifikansnivå, forkaster vi nullhypotesen om at ß = 0. Dette er helt parallelt med vanlig signifikanstesting av koeffisienter i ordinær regresjon.

Siden koeffisientene i logistisk regresjon kun er asymptotisk normalfordelt (dvs. når utvalgsstørrelsen går mot uendelig), er det i prinsippet kun korrekt å bruke z-verdier som kritiske verdier. Aldrich og Nelson (1984:55) anbefaler likevel å bruke Student's t-fordeling fordi denne gir mer konservative hypotesetester. (Siden t-fordelingen har høyere kritiske verdier, skal det mer til for å forkaste null-hypotesen.) Dette argumentet er imidlertid noe merkelig i og med at det forutsettes at en arbeider med så store utvalg at z-fordelingen og t-fordelingen er nærmest identisk. Dersom en arbeider med utvalg hvor fordelingene er relativt ulike, er det grunn til å stille spørsmål om z-testen er en valid test av hypoteser. Da finnes det alternative framgangsmåter som vi kommer tilbake til.

En del statistikkprogrammer rapporterer en skåkalt Wald-statistikk. Ved tohalet hypotesetesting er denne et alternativ til z-testen.[9] Den beregnes rett og slett ved å kvadrere z-operatoren ovenfor, dvs.:

[8] Standardfeilen til et estimat angir i hvilken grad en risikerer at estimatet avviker fra den verdien vi ønsker å estimere (populasjonsverdien).

[9] Tohalet hypotesetest innebærer at en tester hvorvidt en koeffisient er forskjellig fra en oppgitt verdi, f.eks. 0, uavhengig av om den er større eller mindre. Enhalet test innebærer å teste hypoteser om at koeffisienten er enten større eller mindre enn den oppgitte verdien.

$$\text{Wald} = z^2 = \left(\frac{\hat{\beta}}{\text{ASE}} \right)^2$$

Denne testoperatoren er kjikvadratfordelt med frihetsgrader (df) = 1 ved svært store utvalg når ß = 0. Når operatoren overskrider den kritiske verdien, 3,84 ved 5% signifikansnivå, forkastes nullhypotesen.[10] Z-testen og Wald-testen gir identiske resultater.

Z-operatoren og Wald-operatoren fungerer godt for svært store utvalg, men likehood-ratio testen, som vi kommer tilbake til nedenfor, er mer pålitelig ved utvalgsstørrelser som brukes i praksis (Agresti 1996:89). Det er grunn til å være forsiktig i bruken av z-testen/Wald-testen, særlig når utvalget er lite. Hva som er store eller små utvalg i denne forbindelse kommer vi tilbake til nedenfor.

En skal også være klar over en annen svakhet med Wald-testen, nemlig at testen heller ikke er helt pålitelig når den absolutte verdien på den estimerte koeffsienten er svært stor. Hosmer & Lemeshow (1989:17) peker blant annet på undersøkelser som viser at disse to testene ikke sjelden svikter når det gjelder forkaste nullhypotesen (om ingen sammenheng) ved en signifikant koeffisient. I tilfeller med store koeffisienter er den estimerte standardfeilen for stor, slik at en i en del tilfeller beholder nullhypoteser en strengt tatt burde forkastet (SPSS 1993:5). Testen blir med andre ord konservativ etterhvert som størrelsen på koeffisienten (målt i absoluttverdi) øker.

Z-testen og Wald-testen krever store utvalg, men hva er et stort utvalg? Litteraturen gir få rettesnorer her. Long (1997:53) gir likevel noen anbefalinger basert på egne erfaringer. Etter hans mening er det risikabelt å benytte logistisk regresjon på utvalg med mindre enn 100 observasjoner, mens utvalg på over 500 observasjoner skulle være betryggende. Som Long påpeker er det imidlertid kovariatstrukturen i modellen som avgjør hva som kreves av observasjoner.[11] Han lanserer som en tommelfingerregel at det bør være minst 10 observasjoner for hver parameter i modellen, dvs. for hver variabel + konstant. (I vår modell er det tre parametre: konstant, kjønnsparameter og utdanningsparameter). I tillegg kan det være behov for enda flere observasjoner dersom det er lite variasjon i avhengig variabel (dvs. svært skjev fordeling) eller sterk kollinearitet. Enkelte typer regresjonsanalyser, som f.eks. ordinal logistisk regresjon, krever også flere observasjoner. Disse retningslinjene må kun tas som meget upresise antydninger om den nødvendige datamengden ved logistisk regresjon.

På eksempelutskriftene fra nedenfor ser vi at SPSS rapporterer Waldstatistikken mens Stata rapporterer z-testen. SPSS oppgir kjikvadratverdi (Wald), antall frihetsgrader og signifikansnivå. Stata-utskriften oppgir z-verdi og signifikanssannsynlighet.

4.1.2 Likelihood-ratio test

Ett alternativ til testene ovenfor er likelihood-ratio testen. For å beregne den logistiske regresjonsmodellen formuleres det, som vi tidligere har nevnt, en Maximum Likelihood funksjon. Hele beregningsmetoden går ut på å finne de koeffisienter som maksimerer logaritmen av denne funksjonen: log likehood. Dette skjer gjennom en rekke iterasjoner. Interasjonsprosessen starter med en modell som bare inkluderer konstanten, dvs. at koeffisientene for alle de uavhengige variablene er satt til null.[12] Siste iterasjon maksimerer log likelihood for den fulle modellen.

[10] Legg merke til at kritisk verdi på 5 %-nivået for Wald-operatoren er kvadratet av den tilsvarende kritiske verdi for z-operatoren.

[11] Begrepet kovariansstruktur blir forklart nærmere nedenfor. Akkurat nå kan vi nøye oss med å sette likhetstegn mellom kovariatstruktur og regresjonsmodellens kompleksitet, dvs. antall uavhengige variabler og antall verdier på disse variablene. Jo flere variabler og jo flere verdier de har, jo mer kompleks modell.

[12] Under goodness of fit nedenfor skal vi se hvordan en beregner log likelihood for denne enkle utgangsmodellen.

Det er mulig å ta utgangspunkt i denne log likelihoodtesten og teste om det skjer en signifikant endring i log likelihood når vi introduserer en eller flere nye variabler i modellen.[13] Hvis vi kaller log likelihood ved utgangspunktet for L_0 og log likelihood når vi har inkludert en variabel for L_1, kan vi beregne likelihood-ratio testoperatoren ved hjelp av følgende formel:

$$G^2 = -2 (L_0 - L_1) \quad \text{eller} \quad (-2L_0) - (-2L_1)$$

Når vi ganger forskjellen i log likelhood mellom de to modellene med –2 får vi en testoperator hvis sannsynlighetsfordeling er G^2-fordelt. Denne er igjen tilnærmet kjikvadratfordelt med frihetsgrader lik forskjellen i antall uavhengige variabler mellom de to modellene.[14]

Noen eksempler kan illustrere ulike anvendelser av likelihood-ratio testen:

(1) Teste en modell med en variabel i forhold til en modell med bare konstant

Dette tilsvarer å teste om en variabel i en bivariat regresjon er signifikant. Vi kan sette opp følgende hypoteser for hypotesetesting:

H_0: Ingen kjønnsforskjeller med hensyn til lønn. (Vi kan like gjerne bruke fordelingen på avhengig variabel for å predikere utfallet.)

H_A: Forskjell mellom menn og kvinner når det gjelder lønnsnivå.

Nedenfor er det gjengitt en SPSS-utskrift av logistisk regresjon hvor timelønn er den avhengige variabelen og kjønn den uavhengige.

[13] Dette er parallelt med F-testen som benyttes innen OLS for å teste om endringen i forklart varians som følge av å introdusere en eller flere nye uavhengige variabler er signifikant.
[14] Testoperatoren er tilnærmet kjikvadratfordelt når nullhypotesen om ingen sammenheng er korrekt.

Tabell 4.1: Utskrift fra en logistisk regresjonsmodell i SPSS med lønn som avhengig variabel og kjønn som uavhengig. Arbeidslivsundersøkelsen 1993.

```
Dependent Variable..   LONNOMK

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   1904,0501

* Constant is included in the model.

Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..       KJONN     kjønn

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than ,01 percent.

 -2 Log Likelihood     1765,653
 Goodness of Fit       3151,789
 Cox & Snell - R^2         ,043
 Nagelkerke - R^2          ,043

                     Chi-Square    df Significance

 Model                  138,397     1        ,0000
 Block                  138,397     1        ,0000
 Step                   138,397     1        ,0000

Classification Table for LONNOMK
The Cut Value is ,50

                    Predicted
                   ,00    1,00     Percent Correct
                     0  I    1
Observed         +-------+-------+
   ,00      0    I  2869 I     0 I  100,00%
                 +-------+-------+
   1,00     1    I   283 I     0 I     ,00%
                 +-------+-------+
                            Overall  91,02%

---------------------- Variables in the Equation -----------------------

Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

KJONN        -1,7369     ,1730 100,8474     1    ,0000  -,2278     ,1761
Constant     -1,7986     ,0694 671,6970     1    ,0000
```

Omtrent midt på utskriften er det oppgitt tre kjikvadrattester som i denne modellen er identiske. Kjikvadratverdien er 138,397 og den er signifikant med 1 frihetsgrad (fordi det er inkludert 1 variabel i forhold til en modell med bare en konstant). Denne er beregnet ut fra formelen ovenfor. Tidlig i utskriften er det oppgitt – 2 log likelihood når bare konstanten er inkludert (1904,050). Lenger ned finner vi –2 log likelihood etter siste iterasjon (1765,653). Forskjellen mellom disse to, 1904,050-1765,653 gir 138,397.

Testen viser at det å inkludere kjønn i forhold til å utelukkende bruke fordelingen på avhengig variabel til å predikere utfall ga en statistisk signifikant forbedring av likelihood-funksjonen. Vi må følgelig forkaste nullhypotesen om at det ikke er en sammenheng mellom lønn og kjønn.

Denne framgangsmåten kan også brukes til å teste hvorvidt en større modell (flere uavhengige variabler) gir en signifikant forbedring av likelihood-funksjonen i forhold til en modell hvor bare konstanten

er inkludert. Da er det ikke enkeltvariabler som testes, men hele modellen. Antall frihetsgrader ved kjikvadrattesten er lik antall variabler i modellen.

Bruk av likelihoodratiotesten til å sammenlikne en modell med en modell som bare har konstant, tilsvarer F-testen i OLS. Dersom testen er signifikant er fortolkningen ved begge typer tester at minst en av de uavhengige variablene gir et signifikant bidrag til å forklare variasjon i den avhengige variabelen.

Stata gir i prinsippet den samme informasjonen som SPSS, men oppgir log likelihood i stedet for –2 log likelihood (-2LL) slik SPSS gjør. Log likehood for iterasjon 0 tilsvarer log likehoodverdien for en modell som bare inkluderer konstantleddet. Vi ser at log likelihoodverdien er negativ. Ganger vi denne med –2 får vi samme verdi som SPSS oppgir: -2 (-952,02505) = 1904,0501.

Tabell 4.2: Utskrift fra en regresjonsmodell i Stata med lønn som avhengig variabel og kjønn som uavhengig. Arbeidslivsundersøkelsen 1993.

```
. logit lonn_dik kjonn

Iteration 0:   log likelihood = -952.02505
Iteration 1:   log likelihood = -889.46224
Iteration 2:   log likelihood = -882.98664
Iteration 3:   log likelihood = -882.82691
Iteration 4:   log likelihood = -882.82667

Logit estimates                                   Number of obs   =       3152
                                                  LR chi2(1)      =     138.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -882.82667                       Pseudo R2       =     0.0727

------------------------------------------------------------------------------
lonn_dik |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   kjonn |   -1.73702   .1729655    -10.043   0.000      -2.076026   -1.398014
   _cons |  -1.798623    .069399    -25.917   0.000      -1.934643   -1.662603
------------------------------------------------------------------------------
```

Det samme gjelder log likelihood for den fulle modellen. Ganger vi –882,82667 med –2 får vi 1765,65334. Forskjellen mellom 1904,0501 og 1765,6533 er lik 138,3968. Dette er kjikvadratverdien som er oppgitt i begge utskriftene.

(2) Teste om en ny variabel gir et signifikant bidrag.

Vi kan også teste om det å inkludere ytterligere en variabel gir en signifikant forbedring av likelihood-funksjonen. Dette kan regnes ut for hånd, men en kan også få SPSS og Stata til å regne forskjellen. Dersom man kjører en ny modell med både kjønn og utdanning, kan en regne ut om forskjellen i log likelihood mellom de to modellene er statistisk signifikant. I vårt eksempel tester vi følgende hypoteser:

H_0: Ingen utdanningsforskjeller med hensyn til lønn. (Informasjon om utdanning bidrar ikke til å predikere utfallet på avhengig variabel utover informasjon om kjønn)

H_A: Forskjeller i lønnsnivå avhengig i utdanningslengde.

Nedenfor er et utsnitt av SPSS-utskriften for den trivariate analysen:

Tabell 4.3: Utskrift fra en logistisk regresjonsmodell i SPSS med lønn som avhengig variabel og kjønn og utdanning som uavhengige. Arbeidslivsundersøkelsen 1993.

```
 -2 Log Likelihood     1586,990
 Goodness of Fit       3035,467
 Cox & Snell - R^2         ,096
 Nagelkerke - R^2          ,096

                     Chi-Square    df Significance

 Model                  317,061     2        ,0000
 Block                  317,061     2        ,0000
 Step                   317,061     2        ,0000

Classification Table for LONNOMK
The Cut Value is ,50
                    Predicted
                   ,00    1,00     Percent Correct
                     0  I    1
Observed         +-------+-------+
   ,00      0    I  2863 I     6 I   99,79%
                 +-------+-------+
   1,00     1    I   278 I     5 I    1,77%
                 +-------+-------+
                            Overall  90,99%

---------------------- Variables in the Equation -----------------------

Variable            B      S.E.     Wald    df      Sig        R    Exp(B)

KJONN         -1,6026     ,1763  82,6356     1    ,0000  -,2137     ,2014
UTDAAR          ,2987     ,0228 171,5556     1    ,0000   ,3099    1,3482
Constant      -2,9629     ,1279 536,6160     1    ,0000
```

La L_1 betegne log likelihood for den enkleste modellen og la L_2 betegne log likelihood for den utvidede modellen. Vi må nå sammenlikne log likelihood etter siste iterasjon i de to modellene etter følgende formel:

$$\chi^2 \approx G^2 = (-2 \ln L_1) - (-2 \ln L_2) = 1765{,}653 - 1586{,}990 = 178{,}663$$

Denne forskjellen er signifikant med frihetsgrader lik 1. Vi forkaster derfor nullhypotesen og konkluderer med at lønnsnivå henger sammen med utdanningslengde. Antall frihetsgrader er lik 1 fordi vi har inkludert 1 ny variabel i den trivariate modellen. Antall frihetsgrader er lik antall nye variabler vi har i den utvidede modellen.

Både SPSS og Stata tilbyr en måte å gjøre denne testen på hvor en slipper en å regne ut forskjellen selv. I SPSS gjøres dette ved å legge inn variabler i blokker ved hjelp av Block-opsjonen når en bestiller en logistisk regresjon. SPSS vil da teste modellforbedring for hver ny blokk av variabler (som godt kan bestå av bare en variabel). Kijkvadrattesten vil da se slik ut på utskriften:

Tabell 4.4: Kjikvadrattest for forskjellen i log likelihood mellom to modeller ved bruk av Block-opsjonen i SPSS.

```
                     Chi-Square    df Significance

 Model                  317,061     2        ,0000
 Block                  178,664     1        ,0000
 Step                   178,664     1        ,0000
```

Først oppgis kjikvadratverdien for hele modellen, deretter oppgis kjikvadrattesten for siste blokk sammenliknet med forrige.

I Stata kan det samme oppnås, men på en måte som er mer tungvinn enn i SPSS (se Hamilton 1998:236). Først må den enkle modellen kjøres og log likelihoodratiotesten for denne kjøringen lagres. Deretter må den utvidede modellen kjøres. Til slutt må det log likelihood for de to modellene sammenliknes. Kommandoene for å gjøre dette er:

Tabell 4.5: Test av forskjellen i log likelihood mellom to modeller i Stata.

```
. logit lonn_dik kjonn if utdaar~=.                    (1) Kjører enkel modell
. lrtest, saving(0)                                    (2) Lagrer log likelihoodverdi for enkel modell
. logit  lonn_dik kjonn utdaar                         (3) Kjører utvidet modell
. lrtest                                               (4) Ber om log likelihoodratiotest

Logit:  likelihood-ratio test          chi2(-1)    =    -178.66
                                       Prob > chi2 =          .
```

Legg merke til den første kommandoen. Regresjonsanalysen kjøres under betingelse av at det ikke er missing på den variabelen som inkluderes i den utvidete modellen (if utdaar~=.). Grunnen til dette er at antall observasjoner må være likt i de to regresjonskjøringene, ellers er ikke sammenlikningsgrunnlaget likt.[15] Dette er for øvrig et generelt poeng. Det er ytterst viktig i forbindelse med likelihoodratiotesten at utvalgsstørrelsen ikke må endre seg etterhvert som en trekker inn nye variabler. Det gjør nemlig at to modeller ikke er sammenliknbare og en kan heller ikke regne ut forskjeller i log likelihood slik vi har gjort her.

Begge de to testene som gjennomgått ovenfor finnes i de statistikkprogrammer som tilbyr logistisk regresjon. En tredje test, som så langt jeg vet, ikke finnes i noen statistikkpakker er skåretesten (score test) eller Lagrange Multiplier test (LM-test). Denne testen er gjennomgått hos blant annet Hosmer & Lemeshow (1989:17-18) og Long (1997:87-89). Wald-testen, log likelihoodratio testen og skåretesten er alle såkalt asymptotisk ekvivalente. Det innebærer at de alle konvergerer til den samme kjikvadratfordelingen etterhvert som antall observasjoner (N) øker (Long 1997:89). Det er viktig å huske på at alle testene krever relativt store utvalg for at kjikvadratfordelingen skal utgjøre en god tilnærming. Temaet utvalgsstørrelser ble for øvrig behandlet mer inngående under diskusjonen av Waldtesten. Den av disse tre testoperatorene som gjør nytte av mest informasjon er imidlertid log likelihoodratio testen. Denne er også den mest pålitelige av de tre (Agresti 1996:96).

4.2 Goodness of fit

Som forsker ønsker en ofte en mål på hvor «god» en analysemodell er. Et vesentlig spørsmål er imidlertid hva det innebærer at en modell er god, dvs. hva den er god på. Ideelt sett er naturligvis en modell kun god i den forstand den er riktig, dvs. gir uttrykk for de faktiske relasjonene mellom variablene i modellen. Men et slikt mål på såkalt «goodness of fit» forutsetter at vi vet hvordan den riktige modellen ser ut, noe vi naturligvis ikke gjør.

Mål på goodness of fit tar derfor utgangspunkt i det eneste uttrykket for den riktige modellen som vi kan observere, nemlig verdiene på den avhengige variablene. Alle slike mål måler derfor på en eller

[15] Stata vil komme med advarselen «Warning: observations differ» hvis antall observasjoner ikke er likt i de to modellene.

annen måte grad av samsvar mellom de innsamlede dataene og data generert av modellen i form av predikerte eller forventede verdier. Jo større samsvar, jo bedre goodness of fit.

Slike mål er ikke uproblematiske. Hagquist & Stenbeck (1998) referer f.eks. til flere intense diskusjoner omkring nytten og bruken av mål på goodness of fit. Det mest kjente og brukte målet er antakelig R^2 innenfor ordinær regresjonsanalyse. Dette viser blant annet hvor mye av den totale variansen i den avhengige variabelen som en regresjonsmodell gjør rede for. Jo større andel forklart varians, jo bedre modell. Dette er imdlertid avhengig av hva formålet med undersøkelsen er. Dersom vi utelukkende ønsker komme fram til gode prediksjoner av verdier på avhengige variabel er R^2 et egnet mål. I en slik analyse er det ikke vesentlig å operere med en riktig kausalmodell, poenget er om modellen predikerer godt. Dersom siktemålet derimot å teste kausalhypoteser for å avdekke hvor mye konkrete variabler påvirker den avhengige direkte og indirekte, har derimot R^2 liten nytte. En gal modell kan nemlig gi svært høye R^2, noe som bare illustrerer at prediksjon og forklaring er vidt forskjellige siktemål. Et annet problem med R^2 er at målet er følsomt for tilfeldigheter i utvalgets sammensetning.

Achen (1982:67-68) oppsummerer dette på en god måte i sin glitrende monografi om å fortolke og bruke regresjon:

> «In summary, then, any choice among competing regressions is to some extent arbitrary. No choice makes sense outside a theoretical context in which a variety of competing explanations have been tried. A uniform rule for selecting one, such as minimizing R^2 or C_p[16], not only enforces assumptions which one is rarely in complete sympathy, but also violates the nature of the enterprise. By ignoring both prior knowledge and the range og plausible substantive interpretations, routinized procedures subordinate substance to method.»
>
> [...] The second point to be made, and it derives from the first, is that selection of a suitable regression to summarize a data set is an art, not a science. It cannot be reduced to any formal procedure. Perfect regression equations have a manageable number of variables, plausible coefficients, short confidence intervalls, low prediction errors, and a great ease of interpretation»

Når disse harde ord er sagt, gjenstår spørsmålet om logistisk regresjon byr på goodnes of fit mål av typen R^2.

En kan skille mellom absolutte og relative mål for å vurdere hvor god en modell er. De absolutte målene måler samsvar mellom modellprediksjoner og observerte data. Slike mål tar utgangspunkt i fullstendig (mettet) modell som predikerer observasjonene perfekt. De relative målene er ikke relatert til en fullstendig modell, men kan brukes til å finne ut hvilken av to eller flere modeller som passer best til dataene. På tvers av dette skillet kan en også skille mellom deskriptive mål og test statistikker. Den sistnevnte typen har en kjent sannsynlighetsfordeling som gjør det mulig å teste hypoteser om goodness of fit, men den førstnevnte typen kan beskrive forskjeller i tilpasning, men gir ikke mulighet for å teste hypoteser. I følge Hagquist & Stendbeck (1998) er R^2 et deskriptivt goodness of fit mål, dvs. at det ikke kan brukes til å teste om en modell er god eller ikke.

Innen logistisk regresjon kan det vise seg at situasjonen er enda verre. Her finnes det ingen direkte paralleller til R^2, selv om en har forsøkt å konstruere en del pseudomål som er ment å si det samme. Vi skal nedenfor se på ulike tilnærminger til å mål goodness of fit i logistisk regresjon.

[16] Her går det nok litt raskt for Achen. Det skal antakelig stå: «maximizing R^2 or minimizing C_p».

4.2.1 Klassifikasjonstabell

Denne tabellen har som forutsetning at alle enheter i undersøkelsen som får predikert sannsynlighet for å ha verdien 1 på den avhengige variabelen til 0,5 eller høyere settes til verdien 1, mens de som har predikert sannsynlighet under 0,5 settes til verdien 0. En kan dermed sette opp en tabell hvor en sammenlikner observert fordeling med predikert fordeling på den avhengige variabel. Denne tabellen får fire ruter, slik det er vist nedenfor. Denne tabellen kommer automatisk fram i SPSS ved kjøring av logistisk regresjon (se eksempelutskriften ovenfor). I Stata kan en be om en tilsvarende tabell ved å effektuere kommandoen **.lstat** etter å ha gjennomført en logistisk regresjonsanalyse.

Tabell 4.6: Prinsippet for en klassifikasjonstabell.

Observert utfall	Predikert utfall: 0	Predikert utfall: 1
0	Riktig predikert	Feil predikert
1	Feil predikert	Riktig predikert

På bakgrunn av frekevensene i hver enkelt celle kan en beregne prosentandelen som er riktig predikert innenfor hvert observert utfall. Tanken bak en slik tabell er at dersom modellen predikerer gruppemedlemskap perfekt, da har man en sterk indikasjon på at modellen er god.

Hvis vi ser på våre eksempler, så er det slik at den enkleste modellen (utelukkende med kjønnsvariabelen som uavhengig variabel) predikerer 100 prosent riktig for gruppen av observerte 0-verdier, mens den predikerer 0 prosent riktig for gruppen av observerte 1-verdier. Totalt gir dette en riktig prediksjon på 91 prosent. Dette virker imponerende, men vi ville faktisk predikert like godt dersom vi hadde valgt å predikere ut fra modusverdien på den avhengige variabelen (0). Grunnen til dette er at den avhengige variabelen er så sterkt skjevfordelt i utgangspunktet (91 prosent / 9 prosent). Da vil en alltid predikere godt ved å velge modusverdien som prediksjon. Vi ser følgelig at skjevfordelte variabler allerede i utgangspunktet har lite rom for forbedringer av prediksjon.

Vi aner her en av svakhetene ved klassifikasjonstabeller som mål på goodnes of fit. Variabler som er skjevfordelte vil være lette å predikere riktig selv om modellen er ubrukelig som forklaringsmodell. I vårt eksempel er det 9 prosent som har høy lønn. Dersom man ikke har noen forklaringsmodell overhode, vil det lønne seg å tippe at alle enhetene i undersøkelsen har lav lønn. På denne måten vil en tippe riktig i over 90 prosent av tilfellene. Sammenlikning med en avhengig variabel som er fordelt 50/50 vil derfor nødvendigvis bli urettferdig. Spørsmålet er om en klassifikasjonstabell gir noen god og entydig indikasjon på hvor god en modell er når fordelingen på den avhengige variabelen spiller så stor rolle. Hosmer & Lemeshov (1989:147) er inne på det samme:

> *«Classification is sensitive to the relative sizes of the two component groups and will always favor classification into the larger group, a fact that is also independent af the fit of the model.»*

En annen svakhet, som også Hosmer & Lemeshov er inne på, gjelder selv omregningen av sannsynligheter til verdiene 0 og 1. Dette reduserer sannsynligheter målt langs et kontinuum til en todelt variabel. Relativt små forskjeller i sannsynligheter kan her innebære at man bikker over fra den ene gruppen til den andre, men har dette relevans for den prosessen vi prøver å modellere. Etter min mening har det

ikke det. Dette bygger på en individualistisk forestilling om sannsynligheter. Når sannsynligheten for at noe inntreffer blir 0,5 eller høyere vil fenomenet tendere til å inntreffe. En kan nærmest tenke seg en vektskål hvor sannsynligheten for at fenomenet inntreffer veies mot sannsynligheten for at det ikke inntreffer. Vektskålen vil bikke over i den retningen der sannsynligheten veier tyngst.

Jeg har imidlertid i dette notatet lagt vekt på en annen fortolkning av sannsynligheter, nemlig som andeler. Det ligger i dette at fokuset er på andelen enheter innen en gruppe (med like kjennetegn) som utsettes for et fenomen eller utfører en handling. Det er andelen som eksponeres vi måler, ikke individuelle tilbøyeligheter. Dersom f.eks. den predikerte sannsynligheten for å ha verdien 1 på den avhengige variabelen (f.eks. høy lønn) er 0,6 for en definert gruppe (en bestemt kombinasjon av verdier på de uavhengige variablene i modellen), så innebærer det i følgelig klassifikasjonstankegangen at alle i denne gruppen gis verdien 1. Men denne sannsynligheten viser tvert i mot til andelen innenfor denne gruppen som har verdien 1 (60 prosent). Dersom vi gir alle i gruppen verdien 1, så settes samtidig sannsynligheten for å ha høy lønn i denne gruppen til 100 prosent, noe som ikke er i tråd med den predikerte andelen.

En siste kritikk mot klassifikasjonstabeller som mål på tilpasning er at det er lett å tenke seg situasjoner hvor en opererer med en korrekt modell samtidig som klassifiseringstabellen vil gi et dårlig resultat (Hosmer & Lemeshov 1989:146-47). Konklusjonen er derfor at slike tabeller er mest anvendelige siktemålet med analysen eksplisitt er klassifikasjon, eller bør en satse på andre mål på goodnes of fit.

4.2.2 Krysstabelltilnærminger

En annen mulighet til å teste goodness of fit er å konstruere en krysstabell hvor en krysser verdiene på avhengig og de uavhengige variablene. En kan så sammenlikne de observerte frekvensene i tabellen med de predikerte andelene fra regresjonsmodellen. En kjikvadrattest kan så brukes til å teste hvorvidt det er godt samsvar mellom de to tabellene eller ikke. Problemet med denne framgangsmåten er at krysstabellen raskt øker i størrelse når antall uavhengige variabler øker og ikke minst når en inkluderer kontinuerlige variabler. Dette innebærer at det ville være relativt få observasjoner i hver celle, noe som igjen går utover validiteten til kjikvadrattesten og G^2-testen. Et annet problem er at innsamling av mer data ville få tabellen til å øke fordi en ville få flere kombinasjoner av verdier. Tabellen størrelse vil følgelig ikke være fast, men avhengig av datamengden (Agresti 1996:112, Hagquist & Stenbeck 1998:239-240). For å forstå dette må vi gå nærmere inn på begrepene kovariat og kovariatmønstre.

Et meget sentralt begrep innen logistisk regresjon er kovariatmønstre («covariat patterns»). To observasjoner som har identiske verdier på samtlige uavhengige variabler i modellen sies å ha det samme kovariatmønster (StataCorp. 1999:218). I logistisk regresjon er det de observerte og ikke de potensielle kombinasjonene som er vesentlige. For å finne antall kovariater i en modell må en derfor telle opp hvilke faktiske kombinasjoner av verdier på de uavhengige variablene som forekommer i datamaterialet. Nedenfor skal vi vise et eksempel på hva som menes med kovariater.

Det man kan gjøre i en goodness of fit test er å sammenlikne observerte frekvenser og predikerte frekvenser for hvert kovariatmønster. Denne testen kan gjøres i Stata med kommandoen **.lfit** etter å benyttet kommandoene **.logistic** eller **.logit**. Dette er vist nedenfor:

Tabell 4.7: Goodness of fit test i Stata.

```
. lfit

Logistic model for lonn_dik, goodness-of-fit test

       number of observations =      3152
 number of covariate patterns =        13
             Pearson chi2(10) =        19.60
                  Prob > chi2 =         0.0333
```

La oss se nærmere på hvordan denne testen er utregnet og hva den forteller. Først må vi identifisere alle eksisterende kovariatmønstre i datamaterialet. Dette kan blant annet gjøre med kommandoen **.tab**

Tabell 4.8: Kovariatmønstre for en modell med lønnsnivå som avhengig variabel og kjønn og utdanning som uavhengige. Arbeidslivsundersøkelsen 1993.

```
. sort kjonn

. by kjonn: tab utdaar lonn_dik

-> kjonn=     Mann
 utdanning |
      i år |
    utover |   lønn dikotomisert
 grunnskol |       lav        høy |     Total
-----------+----------------------+----------
         0 |       260         14 |       274
         1 |       334         22 |       356
         3 |       547         71 |       618
         5 |       154         43 |       197
         7 |        93         18 |       111
         9 |        68         69 |       137
        12 |         6          5 |        11
-----------+----------------------+----------
     Total |      1462        242 |      1704


-> kjonn=   Kvinne
 utdanning |
      i år |
    utover |   lønn dikotomisert
 grunnskol |       lav        høy |     Total
-----------+----------------------+----------
         0 |       272          1 |       273
         1 |       513          6 |       519
         3 |       270         11 |       281
         5 |       185          7 |       192
         7 |       146         13 |       159
         9 |        21          3 |        24
-----------+----------------------+----------
     Total |      1407         41 |      1448
```

Tabellen viser at det er 7 kombinasjoner mellom ulike verdier på utdanningsvariabelen og verdien mann på kjønnsvariabelen. Tilsvarende er 6 kombinasjoner mellom verdien kvinne på kjønnsvariabelen og verdier på utdanningsvariabelen. For hver av disse observerte kombinasjonene (kovariater) kan vi nå beregne predikerte sannsynligheter ut fra regresjonsmodellen.

Tabell 4.9: Sammenlikning av predikert og observert antall med høy/lav lønn, avhengig av kjønn og utdanning. Arbeidslivsundersøkelsen 1993.

Kovariater		Predikerte sannsynligheter		Predikert antall		Observert antall		Totalt
Kjønn	Utdanning	Lav lønn	Høy lønn	Lav lønn	Høy lønn	Høy lønn	Lav lønn	
Mann	0	0,9509	0,0491	260,54	13,46	260	14	274
	1	0,9349	0,0651	332,82	23,18	334	22	356
	3	0,8876	0,1124	548,55	69,45	547	71	618
	5	0,8129	0,1871	160,15	36,85	154	43	197
	7	0,7051	0,2949	78,27	32,73	93	18	111
	9	0,5681	0,4319	77,83	59,17	68	69	137
	12	0,3493	0,6507	3,84	7,16	6	5	11
Kvinne	0	0,9897	0,0103	270,19	2,81	272	1	273
	1	0,9862	0,0138	511,82	7,18	513	6	519
	3	0,9751	0,0249	274,02	6,98	270	11	281
	5	0,9557	0,0443	183,50	8,50	185	7	192
	7	0,9224	0,0776	146,65	12,35	146	13	159
	9	0,8673	0,1327	20,82	3,18	21	3	24

Vi kan nå ved hjelp av en kjikvadrattest, undersøke hvorvidt de predikerte frekvensene i tabellen ovenfor skiller seg signifikant fra de observerte frekvensene som tabulate-kommandoen ga som resultat. Denne testen kan gjennomføres som en vanlig kjikvadrattest hvor en i hver enkelt celle[17] tar

$$\frac{(\text{Predikert antall - Observert antall})^2}{\text{Predikert antall}}$$

Dersom en summerer disse verdiene i samtlige celler, får en kjikvadratverdien. I dette tilfelle blir verdien 19,6. Antall frihetsgrader for testen er lik

antall kovariater – antall parametre i regresjonsmodellen

I vårt eksempel er det 13 kovariater og 3 parametre i regresjonsmodellen (konstant + de to uavhengige variablene). Dette gir 13 – 3 = 10 frihetsgrader. Med 10 frihetsgrader er den kritiske verdien for kjikvadrattesten 18,307. Vår verdi er høyere enn denne og vi må derfor forkaste nullhupotesen om at det ikke er en forskjell mellom observerte og predikerte verdier. Vi får følgelig ikke støtte for at modellen gir en god tilpasning. Med et lavere signifikansnivå ville vi ha beholdt hypotesen.

Kjikvadrattesten krever store utvalg (Agresti 1996:34). Det er vanlig å anbefale at alle predikerte/forventede frekvenser må være minst 5, men testen kan gi en brukbar tilnærming også når dette ikke er tilfelle (Lillestøl 1991:221). Når en inkluderer flere variabler i modellen, vil antall kovariater øke, noe som også innebærer at en fort får svært få observasjoner pr. kovariat. Validiteten ved kjikvadrattesten blir derfor raskt sviktende når modellene blir mer komplekse. Det er derfor svært begrenset hvor lenge en kan benytte denne testen.

Hosmer & Lemeshow (1989) har derfor lansert et alternativt mål på goodness of fit. Her tar en ikke utgangspunkt i kovariater, men derimot predikerte sannsynligheter på den avhengige variabelen. Testen kan fås i SPSS ved å bestille Hosmer and Lemeshow Goodness-of-Fit Test under Options når en bestiller logistisk regresjon på menyen.

[17] Celle defineres her som hver enkelt kombinasjon av kovariater og verdier på den avhengige variabel. I eksempelet blir det følgelig 13 x 2 = 26 celler.

Tabell 4.10: Hosmer & Lemeshows goodness of fit test. Utskrift fra SPSS.

```
---------- Hosmer and Lemeshow Goodness-of-Fit Test-----------

    LONN_DIK = lav              LONN_DIK = høy

Group   Observed    Expected    Observed    Expected     Total

    1    272,000     270,189       1,000       2,811   273,000
    2    513,000     511,821       6,000       7,179   519,000
    3    270,000     274,014      11,000       6,986   281,000
    4    185,000     183,498       7,000       8,502   192,000
    5    260,000     260,538      14,000      13,462   274,000
    6    334,000     332,817      22,000      23,183   356,000
    7    146,000     146,650      13,000      12,350   159,000
    8    547,000     548,551      71,000      69,449   618,000
    9    342,000     340,908     138,000     139,092   480,000

                     Chi-Square    df Significance

Goodness-of-fit test     4,1943     7        ,7572
------------------------------------------------------------
```

Testen fordeler de predikerte sannsynligheter inn i 10 grupper basert på prosentiler, dvs. at en forsøker å få mest mulig likt antall enheter i gruppene. Ovenfor ser vi at dette ikke har lykkes helt. Det skyldes at vi har relativt få verdikombinasjoner på de avhengige variablene, slik at det blir et mindre antall ulike verdier på sannsynlighetsvariabelen. Derfor blir resultatet 9 grupper som er relativt ulikt inndelt. Dette spiller liten rolle for validiteten av testen.

En kan nå betrakte dette som en krysstabell lik 2*g (2 verdier på avhengig variabel ganger g antall grupper). En kan dermed beregne kjikvadratverdien på egen hånd på vanlig måte. I første celle i tabellen er det observert 272, dersom modellen hadde vært riktig skulle det vært 270,189. Kjikvadrat-verdien i denne cellen er: $(272 - 270{,}189)^2 / 270{,}189 = 0{,}012$. Dersom vi beregner hver av cellene og summerer får vi summen 4,19, det samme som SPSS har kommet til (selvfølgelig). Antall frihetsgrader for denne tabellen er $(g - 2) = 9 - 2 = 7$. Det er m.a.o. ingen grunn til å forkaste nullhypotesen om at observert og forventet fordeling er like. Dette gir god goodness of fit.

Det er interessant å merke seg at de to testene ikke gir samme resultat. Den førstnevnte testen er i så fall den mest pålitelige i og med at Hosmer-Lemeshow testen baserer seg på en indeling av materialet som ikke eksplisitt bygger på den modellen som er lagt til grunn.

4.2.3 R^2

Et mål på goodness of fit som ikke rapporteres av verken SPSS eller Stata tar utgangspunkt i den tradisjonelle formelen for R^2. Dette målet bygger på sammenhengen mellom observerte verdier og predikerte sannsynligheter og rapporterer hvor stor del av variasjonen i den dikotome avhengige variabelen som forklares av de predikerte sannsynlighetene. Menard (1995:23), Long (1997:103) og Agresti (1990:111-112)er blant dem som gjør rede for denne framgangsmåten, men det må også legges til at framgangsmåten er svært lite benyttet. Jeg presenterer det her mer som en introduksjon til pseudo-R^2 som presenteres nærmere nedenfor.

For å beregne R^2 må en lagre de predikerte verdiene fra den logistiske regresjonsanalysen. Dette kan lett gjøres i både SPSS og Stata. Deretter foretar en en bivariat regresjon mellom den opprinnelige avhengige variabelen (0 og 1) og de predikerte sannsynlighetene fra den logisistiske regresjonsmodellen (som varierer mellom 0 og 1). R^2 fra denne regresjonen sier hvor mye av variasjonen i den opprinnelige variabelen som blir forklart av regresjonsmodellen.

Tabell 4.11: Beregning av R^2 for logistisk regresjon med dikotom avhengig variabel i SPSS.

```
. predict p
(option p assumed; Pr(lonn_dik))
(87 missing values generated)

. regress lonn_dik p

  Source |       SS       df       MS                  Number of obs =    3152
---------+------------------------------               F(  1,  3150) =  462.47
   Model |    32.97674     1    32.97674               Prob > F      =  0.0000
Residual |  224.614313  3150  .071306131               R-squared     =  0.1280
---------+------------------------------               Adj R-squared =  0.1277
   Total |  257.591053  3151  .081748985               Root MSE      =  .26703

------------------------------------------------------------------------------
lonn_dik |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       p |   .9939388   .0462189     21.505   0.000       .9033166    1.084561
   _cons |   .0005442   .0063121      0.086   0.931      -.0118321    .0129205
------------------------------------------------------------------------------
```

Ovenfor har vi vist framgangsmåten i Stata. Etter å kjørt logit kommandoen sørger kommandoen **.predict [var.navn]** for at det lagres beregnede sannsynligheter. En bivariat regresjon mellom avhengig variabel og predikerte sannsynligheter gir i vårt ekempel en R^2 på 12,8 %.

Det er flere ting som taler for å bruke R^2 som mål på goodness of fit. Målet er relativt lett å beregne, det tillater direkte sammenlikning mellom logistisk regresjonsmodeller med en rekke mer tradisjonelle teknikker. R^2 er også nyttig når det gjelder å estimere standardiserte, logistiske regresjonskoeffisienter (Menard 1995:23). På den annen side har R^2 også en rekke svakheter. R^2 viser forholdet mellom uforklart og total varians, noe som gir mening i ordinær regresjon bygget på minste kvadraters metode. I logistisk regresjon derimot estimeres koeffisientene ut fra en log likelihoodfunksjon. Den praktiske konsekvensen av dette er at R^2 kan synke når vi inkluderer en ny variabel i modellen selv om den gir et bidrag til log likelihoodfunksjonen. Det er mer riktig å bygge analysen av goodness of fit på kriterier som har vært benyttet for å estimere modellparametrene. Et annet problem med R^2 er at det kun kan benyttes når den avhengige variabelen er dikotom. Ved multinomisk og ordinal logistisk regresjon (se kapittel 6 og 7) kan dette målet på tilpasning ikke brukes.

4.2.4 Pseudo-R^2

Det som kjennetegner de ulike pseudo-R^2 mål som er lansert er at de alle bygger på log likelihoodfunksjonen. Stata og SPSS gir begge mål på goodnes of fit kalt pseudo-R^2, eller likelihood-ratio indeks (LRI). Det finnes flere varianter av dette goodness of fit-målet. Stata rapporterer den enkleste, mens SPSS rapporterer to mer raffinerte mål. Long (1997:104-106) og Aldrich & Nelson (1984:??) presenterer ulike varianter av pseudo-R^2. Jeg vil nedenfor diskutere den enkleste varianten.

Utgangspunktet er log likelihoodfunksjonen som vi har diskutert tidligere. Under beregningen av de logistiske regresjonskoeffisientene maksimeres denne funksjonen. Den laveste verdien log likehoodfunksjonen kan få er i en modell hvor bare konstanten er inkludert. Den høyeste verdien den kan få er dersom vi har en mettet modell, dvs. en modell med like mange parametre som det er observasjoner. Under «ideelle» betingelser vil log likelihoodfunksjonen da ha verdien 0. Vi ser følgelig at vi kan trekke en parallell mellom log likelihoodverdien og variansbegrepet innenfor OLS.

Dersom vi ganger log likelihoodverdien med –2 (=-2LL), kan vi betrakte –2LL for modellen med bare en konstant som parallelt med total varians (total sum of squares, TSS) i avhengig variabel innenfor OLS. Dette er en av de få verdiene en kan beregne selv i forbindelse med en logistisk regresjon. En må da ta utgangspunkt i marginalfordelingen på den avhengige variabelen, som i vårt eksempel er:

Tabell 4.12: Fordeling på den dikotomiserte variabelen lønn. Arbeidslivsundersøkelsen 1993.

	Freq.	Percent	Cum.
lav	2869	91.02	91.02
høy	283	8.98	100.00
Total	3152	100.00	

Formelen for å beregne –2 log likelihood for en modell med bare konstant er:

$$\text{-2 log likelihood} = \text{-2 } (n_1 \ln(n_1) + n_2 \ln(n_2) - n \ln(n))$$

hvor n_1 er antall observasjoner med verdien 0, n_2 er antall observasjoner med verdien 1 og n er antall observasjoner totalt.

Dersom vi tar tallene fra marginalfordelingen ovenfor og setter inn i formelen får vi:

$$\text{-2 log likelihood} = \text{-2 } (n_1 \ln(n_1) + n_2 \ln(n_2) - n \ln(n))$$

$$= \text{-2 } (2869 * \ln(2869) + 283 * \ln(283) - 3152 * \ln(3152))$$

$$= 1904, 05$$

Dette tallet stemmer med det som er oppgitt i SPSS-utskriften og Stata-utskriften (men husk at i det sistnevnte tilfellet må tallet først deles på –2).

-2LL for en konkret modell tilsvarer uforklart varians (error sum of squares, ESS) innen OLS. Det sier noe om avstanden til den mettede (perfekte) modellen. Forskjellen mellom –2LL for en modell med bare en konstant og en konkret modell tilsvarer forklart varians (regression sum of squares, RSS).

Dersom L_0 er log likelihood for en modell med bare konstant og L_1 er log likelihood for en konkret modell, så kan et enkelt mål på pseudo-R^2 være:

$$LRI = 1 - \frac{\ln L_1}{\ln L_0} \qquad \text{eller} \qquad LRI = 1 - \frac{-2 \ln L_1}{-2 \ln L_0}$$

Dette målet en intuitiv appell fordi det varierer mellom 0 og 1. Når LRI har verdien 0 bidrar ikke modellen med noen forbedring og alle koeffisientene har verdien 0. I motsetning til R^2 i ordinær regresjon er det ikke mulig å få pseudo-R^2 til anta verdien 1, men en kan komme ganske nær.[18] Dessverre har ikke dette målet noen naturlig fortolkning (Greene 1993:651; Hamilton 1998:231) av typen forklart varians i OLS. Pseudo-R^2 gir likevel mulighet til å sammenlikne hvor godt ulike modeller for samme avhengige variabel i samme utvalg passer.

[18] Grunnen til at pseudo-R^2 ikke kan bli 1 er at en ikke kan ha en perfekt modell ved logistisk regresjon, dvs. en modell som predikerer verdiene på den avhengige variabelen (0 og 1) perfekt. Vi kan forstå dette intuitivt ved å tenke på at logistisk regresjon bygger på forutsetningen om at sammenhengen mellom den avhengige variablene og de uavhengige kan beskrives ved hjelp av S-kurver som nærmer seg 0 og 1 asymptotisk, dvs. aldri helt når disse punktene. En perfekt modell forutsetter at en kan nå punktene 0 og 1 og derfor bryter modellen sammen: matrisen med estimatene «eksploderer» i løpet av iterasjonsprosessen, som det heter i fagliteraturen.

Anvendt på vårt eksempel, finner vi at pseudo-R^2 for den bivariate regresjonsanalysen er:

$$\text{Pseudo-}R^2 = \text{LRI} = 1 - (1765{,}653/1904{,}050) = 1 - 0{,}927 = 0{,}073$$

For den trivariate analysen er pseudo-R^2:

$$\text{Pseudo-}R^2 = \text{LRI} = 1 - (1586{,}99/1904{,}050) = 1 - 0{,}833 = 0{,}167$$

Den siste modellen «passer» i følge dette målet bedre enn den første. Det er for øvrig noe vi allerede fant ut i forbindelse med signifikanstestene, i og med at variabelen utdanning ga en signifikant forbedring av log likelihood når den ble inkludert i modell hvor kjønn allerede var tatt med som uavhengig variabel.

Vi må imidlertid vokte oss vel for å fortolke tallene ovenfor som prosent forklart varians eller liknende. Dette gjelder også de mer raffinerte variantene.

4.2.5 Relativt mål på goodness of fit

Vi har ovenfor lansert log likelihoodratio som et signifikansmål. Men det kan også betraktes som et relativt mål på goodness of fit. Grunnen er at vi direkte kan ta forskjellen mellom –2 log likelihood-verdiene for to ulike modeller og teste hvorvidt den utvidete modellen gir en signifikant forbedring av log likelihood enn den enkle modellen. Denne forskjellen er kjikvadratfordelt gitt at nullhypotesen er sann. Framgangsmåten er beskrevet ovenfor. Vi må bare huske at antall frihetsgrader er lik forskjellen i antall variabler mellom de to modellene. Til praktisk modelleringsarbeid er dette ofte et mer anvendelig verktøy enn absolutte mål på goodness of fit.

4.2.6 Oppsummering

Mål på goodness of fit i logistisk regresjon er på mange måter ikke noe enklere enn i OLS. Særlig sårt for menge er nok mangelen på et surrogat for R^2 (som i seg selv på mange måter er et goodness-of-fit surrogat). Samtidig er egentlig ikke mulighetene mindre i logistisk regresjon. Poenget er kanskje at en må være mer bevisst hvilket mål en velger og se dette i forhold til siktemålet med undersøkelsen. Nedenfor er noen antydninger til retningslinjer når det gjelder bruk av mål på goodness of fit:

1. Man kan gjerne bruke goodness of fit i analysearbeidet, men ikke slavisk for å avgjøre hvilken modell som er best. Den beste modellen er ikke nødvendigvis den som predikerer best i datamaterialet.
2. Hva som er en god modell avgjøres ved å bygge regresjonsmodeller ut fra teoretiske betraktninger, ikke ved å inkludere/ekskludere variabler ut fra rent statistiske vurderinger.
3. Legg vekt på signifikanstester og fortolkbarhet som kriterier for å vurdere ulike modellers nytte.
4. Skal man rapportere mål på goodness of fit (fordi det kreves i internasjonale tidsskrifter e.l.), bør en velge log likelihoodratio test og/eller Hosmer-Lemeshow testen. Ulike pseudo-R^2 bør antakelig unngås. Det er for øvrig mer og mer vanlig at mål på goodness of fit ikke rapporteres i forbindelse med logistisk regresjon. Verden går framover.

4.3 En advarsel

Noe av problemet med logistisk regresjon (og dikotome avhengige variabler generelt) er at en ikke har direkte tilgang til sannsynlighetene. Det er derfor ikke mulig å kontrollere ens funn mot det faktisk mønsteret i datamaterialet. Forholdet mellom regresjonsresultater og data, dvs. den mekanismen som produserer regresjonsresultatene blir en svart boks. Dette gjelder nok også ordinær regresjon (spesielt sannsynlighetsregresjon), men antakelig i enda større grad logistisk regresjon. Selv om vi har fått sig-

nifikante resultater betyr ikke det at vi nødvendigvis har en god modell. En del bruk av logistisk regresjon bærer nok noe preg av at en tror at logistisk regresjon løses alle problemer. Tabellen nedenfor kan være en tankevekker i så måte.

I tabellen har vi beregnet den faktiske andelen som har høy lønn avhengig av kjønn og utdanningslengde og sammenstilt dette med den predikerte andelen. Som tabellen viser er ikke samsvaret mellom observert og predikert like godt i alle kategorier. For kvinner predikerer modellen ganske godt, men vi huske på at vi strengt tatt ikke kan predikere utover 9 års utdannelse fordi vi ikke har data i dette området. For menn ser vi at samsvaret er tilfredsstillende fram til 5 års utdannelse, men at det etter dette er til dels store forskjeller mellom predikert og observert frekvens. For gruppen som har 12 års utdannelse etter grunnskolen vil vi faktisk predikere feil med omtrent 20 prosentpoeng.

Tabell 4.13: Observerte og predikerte sannsynligheter for å ha høy lønn avhengig av utdanning og kjønn. Arbeidslivsundersøkelsen 1993.

	Kvinner		*Menn*	
Utdanning	Observert	Predikert	Observert	Predikert
0	0,4 %	1,0 %	5,1 %	4,9 %
1	1,2 %	1,4 %	6,2 %	6,5 %
2	-	1,9 %	-	8,6 %
3	3,9 %	2,5 %	11,5 %	11,2 %
4	-	3,3 %	-	14,6 %
5	3,7 %	4,4 %	21,8 %	18,7 %
6	-	5,9 %	-	23,7 %
7	8,2 %	7,8 %	16,2 %	29,5 %
8	-	10,2 %	-	36,0 %
9	12,5 %	13,3 %	50,4 %	43,2 %
10	-	17,1 %	-	50,6 %
11	-	21,8 %	-	58,0 %
12	-	27,3 %	45,5 %	65,1 %

Det er derfor viktig å huske at også logistisk regresjon har forutsetninger om formen på sammenhengen mellom avhengig og uavhengig variabel. Modellen forutsetter at disse sammenhengene kan beskrives med en logistisk kurve som er beskrevet tidligere. Denne forutsetningen kan imidlertid være like urealistisk som forutsetningen om sammenhengen kan beskrives med en rett linje slik OLS forutsetter. Det kan derfor være nyttig å sjekke modellens prediksjoner mot observert andel slik det er gjort i tabellen ovenfor. Problemet er at en slik kontroll er vanskelig dersom antall variabler er forholdsvis stort.

4.4 Forutsetninger for bruk av logistisk regresjon

Når vi allerede er inne på forutsetningene bak logistisk regresjon kan være på sin plass å nevne de viktigste begrensningene. På samme måte som OLS, bygger også logistisk regresjon på visse forutsetninger. Mange av dem er identiske med dem man finner i OLS (se f.eks. Lewis-Beck 1980 eller Gujarati 1988:52-60).

Vi har allerede vært inne på en av forutsetningene ovenfor, nemlig at en ikke har tilnærmet perfekt modellspesifikasjon. Da bryter nemlig modellensammen. Hvis så er tilfelle, kan en imidlertid bruke ordinær regresjon, dvs. lineær sannsynlighetsregresjon. I det tilfelle vil nemlig problemene med linær sannsynlighetsregresjon være uten betydning.

Ellers gir Aldrich & Nelson følgende oversikt over forutsetninger som må være tilfredsstillt:

1. Den avhengige variabelen kan anta to verdier, f.eks. 0 og 1. Den primære interessen er å få estimert sannsynligheter for å ha verdien 1 på den avhengige variabelen eller andelen som har verdien 1.
2. Sannsynligheten eller andelen antas å være avhengig av et sett av uavhengige variabler: sammenhengen antas å kunne spesifiseres på følgende måte (dvs. følger den logistiske kurven):

$$P(Y=1) = \frac{e^{(b_0+b_1x_1+\ldots+b_nx_n+e)}}{\left(1+e^{(b_0+b_1x_1+\ldots+b_nx_n+e)}\right)}$$

3. Observasjonene av den avhengige variabelen forutsettes å være statistisk uavhengige av hverandre, dvs. vi har et tilfeldig utvalg av enheter.
4. Vi må ikke ha sterk multikollinearitet, dvs. at eksisterer sterk lineær sammenheng mellom to eller flere av de uavhengige variablene i modellen. Dette er helt parallelt med forutsetningen bak OLS.

5 Innledende om logistisk regresjon når avhengig variabel har mer enn to verdier

5.1 Nominal- og ordinalvariabler

Alle utvidelser av den logistiske regresjonsmodellen bygger på grunnprinsippene i binomisk logisk regresjon, dvs. logistisk regresjon med dikotom avhengig variabel. Binomisk logistisk regresjon kan faktisk betraktes som et spesialtilfelle av multinomisk og ordinal logisistisk regresjon.

Utgangspunktet for utvidelsene av den binomiske logistiske regresjonsmodellen er at vi nå har en avhengig variabel med mer enn to verdier, men som samtidig ikke kan sies å være på forholdstallsnivå. Det første tilfellet vi tar opp er når den avhengige variabelen på nominalnivå, dvs. at verdiene på variabelen markerer utelukkende at vi har gjensidig utelukkende kategorier. Det gir ikke mening å rangordne verdiene. Det andre tilfellet gjelder variabler hvor verdiene kan rangordnes, men hvor det ikke gir mening å snakke om avstanden mellom kategoriene.[19]

For begge disse tilfellene finnes det analysemodeller innen logistisk regresjon. De bygger begge på samme prinsippet, nemlig at den avhengige variabelen «gjøres om» til et sett av dikotome variabler før regresjonsanalysen gjennomføres. Nedenfor gir jeg en innføring i ulike måter å gjøre en variabel om til et sett av dikotomier.

5.2 Håndtering av nominal- og ordinalvariabler – koding til dummyvariabler

5.2.1 Som uavhengige variabler i ordinær og logistisk regresjon

Koding av variabler til dummyvariabler er en velkjent måte å behandle uavhengige variabler på i regresjon. Dersom vi har en variabel på nominalnivå kan denne gjøres om til et sett av dikotome variabler. Antallet dummyvariabler må imidlertid være en mindre enn antall verdier på den opprinnelige variabelen, ellers får vi perfekt kolinearitet.

Dersom vi har en variabel for bosted med følgende fem verdier: øst, sør, vest, midt og nord, kan vi gjøre om variabelen til følgende fire dummyvariabler:

- *sør* med verdien 1 dersom respondenten bor i dette området og verdien 0 ellers
- *vest* med verdien 1 dersom respondenten bor i dette området og verdien 0 ellers
- *midt* med verdien 1 dersom respondenten bor i dette området og verdien 0 ellers
- *nord* med verdien 1 dersom respondenten bor i dette området og verdien 0 ellers

Tar vi disse variablene med som uavhengige variabler i en regresjonsanalyse, vil de koeffisientene vi får vise forskjellen i avhengig variabel mellom det aktuelle området (sør, vest, midt eller nord) og det området som er utelatt (referansekategorien: øst).

5.2.2 Som avhengige variabler i logistisk regresjon

Logistisk regresjon bygger på odds, dvs. forholdet mellom antallet som er i en kategori og antallet som ikke er i denne kategorien. For å bruke logistisk regresjon på variabler med mer enn to verdier må disse variablene gjøres om til et sett av dikotomier, slik at en kan konstruere odds mellom

[19] Hva som kjennetegner ulike målenivåe er beskrevet mer utførlig i avsnitt 2.1.

sannsynligheten for å ha den ene verdien og sannsynligheten for å ha den andre verdien. Nominal logistisk regresjon og ordinal logistisk regresjon omdanner den opprinnelige variabelen på forskjellige måter.

Ved *nominal logistisk regresjon* er den avhengige variabelen på nominalnivå. Framgangsmåte er å velge en referanseverdi (baseline-category). Antallet respondenter i de andre verdiene ses i forhold til antallet respondenter i referanseverdien. En annen måte å forestille seg dette på er at vi konstruerer et sett dikotome variabler som gis verdien 0 når enhetene tilhører referanseverdien og verdien 1 når de tilhører den verdien vi vil sammenlikne referanseverdien med. De andre verdiene på den opprinnelige variabelen ser vi bort fra. Dersom vi bruker eksempelet ovenfor, og fortsatt velger øst som referanseverdi, får vi følgende dikotome variabler:

- *sør* med verdien 1 dersom respondenten bor på Sørlandet og verdien 0 dersom vedkommende bor på Østlandet
- *vest* med verdien 1 dersom respondenten bor på Vestlandet og verdien 0 dersom vedkommende bor på Østlandet
- *midt* med verdien 1 dersom respondenten bor i Midt-Norge og verdien 0 dersom vedkommende bor på Østlandet
- *nord* med verdien 1 dersom respondenten bor I Nord-Norge og verdien 0 dersom vedkommende bor på Østlandet

Vi får alltid en dummyvariabel mindre enn antall verdier på den opprinnelige variabelen. Den videre framgangsmåten er beskrevet nærmere i kapittel 6.

En annen måte å kode om variabler på, er *kumulativt*. Denne benyttes i rangert (ordinal) logistisk regresjon, metoden for å håndtere avhengige variabler på ordinalnivå. Kumulativ koding innebærer at en lager dummyvariabler hvor alle enheter som befinner seg i en kategori eller lavere på den opprinnelige gis en felles verdi (0). De enhetene som har høyere verdier gis en annen felles verdi (1). Det sier seg selv at variabelen må kunne rangeres, dvs. være på ordinalnivå, for at denne framgangsmåten skal kunne benyttes. Også ved denne typen omkoding må antallet dummyvariabler være en mindre enn antall verdier på den opprinnelige variabelen.

La oss ta eksempel med en variabel for sosial status med fem verdier: lavere arbeiderklasse, høyere arbeiderklasse, lavere middelklasse, høyere middelklasse og overklasse. Denne variabelen kan kodes om til fire kumulative dummyvariabler:

- *høyere arbeiderklasse, middelklasse eller overklasse*, med verdien 1 for alle som tilhører kategorien høyere arbeiderklasse eller høyere og verdien 0 for dem som er i kategorien lavere arbeiderklasse
- *middel- eller overklasse*, med verdien 1 for alle som tilhører en av funksjonærkategoriene og verdien 0 for dem som tilhører en av arbeiderkategoriene
- *høyere middelklasse eller overklasse*, med verdien for alle som tilhører høyere middelklasse eller overklasse og verdien 0 for alle som tilhører arbeiderklasse eller lavere middelklasse
- *overklasse*, med verdien 1 for alle som tilhører overklasse og verdien 0 for alle som tilhører arbeider klasse eller middelklasse

Den femte dummyvariabelen er unødvendig fordi vi ser at den vil ha verdien 0 eller 1 for alle respondenter. Denne måten å kode om den avhengige variabelen på benyttes i ordinal logistisk regresjon. Vi har ikke noen referansegruppe, men dummyvariablene relaterer seg til hverandre ved å vise kumulative tall. Antall respondenter med verdien 1 på variabelen *overklasse* må være minst like høyt som antall spondenter med verdien 1 på variabelen *høyere middelklasse eller overklasse*, etc.

6 Multinomisk logistisk regresjon

En rekke problemstillinger innebærer analyse av variabler som har mer enn to verdier og som samtidig ikke kan rangordnes. Slike variabler er på nominalnivå. Dette er det laveste av målenivåene og det er begrenset hvilke regneoperasjoner som kan utføres på denne typen variabler (Hellevik 1991:155-56). Eksempler på nominalvariabler kan være:

- yrke
- bosted
- fritidsinteresse
- nasjonalitet/etnisk opprinnelse
- partitilhørighet uten å rangere langs høyre-/venstredimensjon e.l.

Logistisk regresjon bygger imidlertid på oddsratioen, dvs. forholdet mellom sannsynligheten for at en egenskap er tilstede og sannsynligheten for at den ikke er tilstede. For å kunne gjennomføre logistisk regresjon med en avhengig variabel med mer enn to verdier, må en derfor se på forholdet mellom to og to kategorier. En kan si at den avhengige variabelen gjøres om til et sett av dikotome variabler som så analyseres slik det ble framstilt i kapittel 2. En får med andre ord like mange regresjonslikninger som det er dummyvariabler. Dette settet av likninger må løses simultant.

Ved multinomisk regresjon konstrueres det oddsratioer hvor sannsynligheten for å ha en verdi på den avhengige variabelen ses i forhold til sannsynligheten for å ha en referanseverdi. Det innebærer at det blir en dummyvariabel mindre enn det er verdier på den opprinnelige variabelen. Den avhengige variabelen viser da oddsen for å ha en bestemt verdi på variabelen i forhold til en referanseverdi.

La oss ta som eksempel *hva forbrukerne hovedsakelig legger vekt på ved kjøp av produkter*. Denne variabelen kan f.eks. ha verdiene: pris, kvalitet/effektivitet tilgjengelighet og miljø/helse. Dersom vi er interessert i å undersøke hvilke faktorer som har betydning for hva forbrukerne legger vekt på, kan dette gjøres ved hjelp av nominal logistisk regresjon. Vi må først velge en referanseverdi som vi vil sammenlikne de andre verdiene med. En naturlig referanseverdi kunne f.eks. være pris. Det innebærer at vi ønsker å se hva som kjennetegner:

- kvalitets-/effektivitetsorienterte forbrukerne i forhold til prisorientert forbrukere
- tilgjengelighetsorienterte forbrukere opp i forhold til prisorientert forbrukere
- miljø-/helseorienterte forbrukere i forhold til prisorienterte forbrukere

Vi får tre sammenlikninger ut fra de fire verdiene på den opprinnelige variabelen. Dette kan også formuleres ved hjelp av følgende likninger:

$$L_1 = \ln\left(\frac{P(Y = \text{kvalitet})}{1 - P(Y = \text{pris})}\right) = b_{01} + b_{11}x_1 + ... + b_{n1}x_n + e$$

$$L_2 = \ln\left(\frac{P(Y = \text{tilgjeng.})}{1 - P(Y = \text{pris})}\right) = b_{02} + b_{12}x_1 + ... + b_{n2}x_n + e$$

$$L_3 = \ln\left(\frac{P(Y = \text{miljø})}{1 - P(Y = \text{pris})}\right) = b_{03} + b_{13}x_1 + ... + b_{n3}x_n + e$$

Vi ser at b'ene i likningene er markert med to numre. Det første viser hvilken variabel koeffisienten gjelder. Det andre viser hvilken likning koeffisienten gjelder. Grunnen til at det er behov for det siste tallet er at vi når får ulike koeffisienter for hver av variablene, avhengig av hvilke verdier vi sammenlikner.

Substansielt er ikke dette så vanskelig å forstå. Vi kan tenke oss kjønn som en uavhengig variabel. Det er ikke rimelig å tro at kjønn spiller den samme rolle i sammenlikningen mellom kvalitetsorienterte og prisorientert forbrukere som i sammenlikningen mellom miljø-/helseorienterteog prisorienterte forbrukere. Vi får med andre ord flere kjønnskoeffisienter.

Vi ser også at valget av referanseverdi er svært viktig. Det er viktig å velge en verdi som det er teoretisk interessant å sammenlikne med de andre verdiene. Dette er blant annet avhengig av problemstilling. Dersom problemstillingen var rettet mot miljø og helse, ville det være mer naturlig å velge denne kategorien som sammenlikningsverdi for de andre.

Når en gjennomfører multinomisk logistisk regresjon vil en som nevnt få flere estimater for hver variabel, en for hver sammenlikning. En kan derfor spørre om det ikke heller ville være like greit å gjennomføre separate logistiske regresjoner, hvor en sammenlikner to og to verdier på den avhengige variabelen. Det er to grunner til at dette ikke er å anbefale. For det første sikrer multinomisk logistisk regresjon at summen av sannsynlighetene for å falle inn under en av kategoriene på den avhengige variabelen alltid blir 1. Det blir ikke nødvendigvis tilfelle når en kjører separate regresjoner. For det andre gir multinomisk logistisk regresjon mer effektive estimater, fordi standardfeilene til koeffisientene blir mindre (Agresti 1996:206).

Ved estimering av koeffisientene tar multinomisk logistisk utgangspunkt i de likningene som ble vist ovenfor. Modellen forsøker å finne en løsning som medfører at det observerte materialet, dvs. kombinasjonene av verdier på den avhengige og de uavhengige variablene blir mest mulig sannsynlig. Dette foregår ved å finne de løsningen som gir den høyeste log likelihoodverdien. På samme måte som ved binær logistisk regresjon, må løsningen finnes numerisk, dvs. ved hjelp av interasjoner.

Siden framgangsmåten ved multinomisk regresjon i prinsippet følger samme framgangsmåte som binær logistisk regresjon, er det ikke noe å tilføye når det gjelder bruk av signifikanstester og mål på goodnes of fit. Det eneste skillet er at vi nå får flere z-tester for hver av de uavhengige variablene.

Når det gjelder fortolkningen av koeffisienter er også det som ble sagt i kapittel to gyldig. Det eneste nye er også her at vi får flere koeffisienter for hver variabel og at fortolkningen av oddsene hele tiden er forholdet mellom forekomesten av den aktuelle verdi og forekomsten av referanseverdien.

For å illustrere bruk og fortolkning av nominal logistisk regresjon, gjennomgår jeg to eksempler nedenfor.

6.1 Eksempel 1: Fagforeningsmedlemskap og kjønn

6.1.1 Innledende krysstabell: odds og oddsratioer

Den avhengige variabelen i dette eksempelet er fagforeningsmedlemskap med fem verdier: ikke medlem i fagforening, medlem i LO, medlem i YS, medlem i AF og medlem av frittstående fagforening. Tabellen nedenfor viser fordelingen av fagforeningsmedlemskap avhengig av kjønn.

Fagforeningsmedlemskap	*Menn*	*Kvinner*	*Total*
Ikke medlem	35	29	33
Medlem LO	37	32	35
Medlem YS	8	19	13
Medlem AF	14	10	12
Medlem frittstående	6	10	7
Total	100 (1735)	100 (1474)	100 (3209)

Den største forskjellen i tabellen er at andelen som er medlem av YS er større for kvinner enn for menn. Det er en noe større andel menn som ikke organisert, samt organisert i LO og AF. Dersom vi velger *ikke medlem* som referansekategori på den avhengige variabelen, kan vi beregne odds for de ulike verdiene i forhold til referanseverdien. Disse er vist i neste tabell.

Fagforeningsmedlemskap	*Odds for menn*	*Odds for kvinner*	*Oddsratio kvinner/menn*
Ikke medlem	-	-	-
Medlem LO	37/35 = 1,06	32/29 = 1,10	1,10/1,06 = 1,04
Medlem YS	8/35 = 0,23	19/29 = 0,66	0,66/0,23 = 2,87
Medlem AF	14/35 = 0,40	10/29 = 0,34	0,34/0,40 = 0,85
Medlem frittstående	6/35 = 0,17	10/29 = 0,34	0,34/0,17 = 2,00

Både for menn og kvinner er oddsen for å være medlem i LO høyere enn 1. Det innebærer at det er noe mer sannsynlig at både menn og kvinner er medlem av LO enn at de ikke er medlem av noen fagorganisasjon. De andre oddsene er mindre enn 1. Det innebærer at medlemskap i YS, AF eller frittstående fagforening er mindre sannsynlig enn å ikke være medlem.

Oddsratioene viser oddsen for kvinner i forhold til oddsen for menn. Er oddsratioen større enn 1, er sjansen for å tilhøre kategorien størst for kvinner, og omvendt dersom oddsratioen er mindre enn 1. Oddsen for å tilhøre LO er omtrent like stor for både menn og kvinner. Oddsen for å være medlem av YS og frittstående er størst for kvinner, mens oddsen for å tilhøre AF er størst for menn (men forskjellen er ikke stor).

6.1.2 Multinomisk logistisk regresjon – tolkning av koeffisienter

Multinomisk regresjon kan ikke gjøres i SPSS. Vi har derfor benyttet kommandoen **.mlogit** i Stata. Resultatet er vist i boksen nedenfor.

Vi ser at vi nå får fire sette med koeffisienter, en for hver av de parvise sammenlikningene. Den første viser kjønnseffekten når det gjelder sammenlikning av sannsynligheten for å være medlem av LO i forhold til å ikke være medlem av fagforening. Vi ser at kjønnskoeffisienten ikke er signifikant. Det innebærer at oddsen for å medlem av LO snarere enn å ikke være medlem av fagforening, ikke er signifikant forskjellig for menn og kvinner. Det er i tråd med tabellen ovenfor, hvor vi fant at oddsratioen lå svært nær 1.

Kjønnskoeffisienten i den andre parvise sammenlikningen er positiv og statistisk signifikant. Det innebærer at sannsynligheten for å være medlem av fagforening i forhold til sannsynligheten for ikke å være medlem av fagforening overhode, er større for kvinner enn for menn.

```
. mlogit fagfor kjonn, b(0)

Iteration 0:  Log Likelihood = -4649.568
Iteration 1:  Log Likelihood =-4589.0315
Iteration 2:  Log Likelihood =-4587.5702
Iteration 3:  Log Likelihood =-4587.5687

Multinomial regression                                  Number of obs =   3209
                                                        chi2(4)       = 124.00
                                                        Prob > chi2   = 0.0000
Log Likelihood = -4587.5687                             Pseudo R2     = 0.0133

------------------------------------------------------------------------------
  fagfor |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Medlem L |
   kjonn |   .0363569   .0873539      0.416   0.677      -.1348537    .2075674
   _cons |   .0460034   .0563362      0.817   0.414      -.0644135    .1564203
---------+--------------------------------------------------------------------
Medlem Y |
   kjonn |   1.086391     .12183      8.917   0.000       .8476086    1.325173
   _cons |  -1.510592   .0947432    -15.944   0.000      -1.696285   -1.324899
---------+--------------------------------------------------------------------
Medlem A |
   kjonn |  -.1304458   .1218593     -1.070   0.284      -.3692857    .1083941
   _cons |    -.93845   .0759787    -12.351   0.000      -1.087365   -.7895346
---------+--------------------------------------------------------------------
Medlem f |
   kjonn |   .7591616   .1454283      5.220   0.000       .4741273    1.044196
   _cons |  -1.848536   .1092367    -16.922   0.000      -2.062636   -1.634436
------------------------------------------------------------------------------
(Outcome fagfor==Ikke med is the comparison group)
```

I den tredje parvise sammenlikningen er kjønnskoeffisienten negativ, men ikke signifikant. Det innebærer at sjansen for å være medlem av AF ikke er signifikant forskjellig for menn og kvinner. Det er også i tråd med tabellanalysen.

I den siste sammenlikningen er kjønn igjen positiv og signifikant. Det innebærer at oddsen for å være medlem av frittstående fagorganisasjon snarere enn ikke å være medlem av noen fagforening, er større for kvinner enn for menn.

Oddsratioene kan for øvrig beregnes ved hjelp av følgende formel (jf. avsnitt 3.4.2):

$$\theta = e^{b} = e^{0,04} = 1,04$$
$$\theta = e^{b} = e^{1,09} = 2,97$$
$$\theta = e^{b} = e^{-0,13} = 0,88$$
$$\theta = e^{b} = e^{0,76} = 2,14$$

Vi ser at de estimerte oddsratioene stemmer svært godt overens med oddsratioene i tabellanalysen. Det behøver ikke nødvendigvis være tilfelle.

6.1.3 Beregning av sannsynligheter

Det er mulig å regne ut sannsynlighetene for å tilhøre de ulike kategoriene. Beregningene er ikke vanskelige, men kan være noe tidkrevende.

(1) Først må *samtlige odds* beregnes for den kombinasjonen av verdier på de uavhengige variablene vi ønsker å beregne sannsynlighet for.

Hvis vi f.eks. ønsker å beregne sannsynligheten for å være medlem av AF for kvinner, må vi beregne oddsen for å være medlem av LO, YS, AF og frittstående for kvinner:

$$\theta_{LO} = e^{(0,05 + (1 \times 0,04))} = e^{(0,09)} = 1,09$$

$$\theta_{YS} = e^{(-1,51 + (1 \times 1,09))} = e^{(-0,42)} = 0,66$$

$$\theta_{AF} = e^{(-0,94 + (1 \times 0,13))} = e^{(-0,81)} = 0,44$$

$$\theta_{fritt} = e^{(-1,85 + (1 \times 0,76))} = e^{(-1,09)} = 0,34$$

(2) Oddsen for den aktuelle verdien på avhengig variabel *deles på summen av 1 og samtlige odds*.

p(Y=YS | X=kvinner) = 0,66 / (1 + 1,09 + 0,66 + 0,44 + 0,34) = 0,66 / 3,53 = 0,19

Vi kan også regne ut sannsynlighetene for de andre verdiene på avhengig variabel:

p(Y=LO | X=kvinner) = 1,09 / (1 + 1,09 + 0,66 + 0,44 + 0,34) = 1,09 / 3,53 = 0,31
p(Y=AF | X=kvinner) = 0,44 / (1 + 1,09 + 0,66 + 0,44 + 0,34) = 0,44 / 3,53 = 0,12
p(Y=fritt | X=kvinner) = 0,34 / (1 + 1,09 + 0,66 + 0,44 + 0,34) = 0,34 / 3,53 = 0,10

For referansekategorien er formelen noe annerledes. Her deles 1 på summen av 1 og samtlige odds.

p(Y=ingen | X=kvinner) = 1 / (1 + 1,09 + 0,66 + 0,44 + 0,34) = 0,66 / 3,53 = 0,28

Legger vi sammen de estimerte sannsynlighetene (0,19+0,31+0,12+0,10+0,28) får si summen 1. Det skal alltid være tilfelle. Vi ser også at de predikerte sannsynlighetene/andelene ikke avviker mye fra de faktiske andelene i den opprinnelige krysstabellen.

6.1.4 Signifikanstester og «goodnes of fit»

Det er svært liten forskjell på signifikanstester og mål på goodnes of fit i binomisk og multinomisk logistisk regresjon. Den eneste forskjellen er at multinomisk regresjon gir flere z-tester for hver av de uavhengige variablene, en for hvert sett av regresjonskoeffisienter. Det er, som nevnt tidligere, viktig å huske at de forutsetter et relativt stort utvalg for å være pålitelige. Dersom en er i tvil om utvalgsstørrelsen er tilstrekkelig, bør en heller benytte likelihoodratio-testen. Denne siste testen tester imidlertid kun om en variabel som helhet bidrar til en forbedring av log likelihood, ikke om variabelen er statistisk signifikant i den enkelte delmodell som kommer fram i utskriften.

Når det gjelder goodness of fit er det også visse forskjeller i forhold til logistisk regresjon med dikotom avhengig variabel. Klassifikasjonstabeller, krysstabelltilnærminger og R^2 kan ikke lenger benyttes som mål på tilpasning. Det er kun pseudo-R^2 og mål som bygger på log likelihood som kan benyttes.

I eksempelet ovenfor kan vi teste forskjellen i loglikelihood mellom den aktuelle regresjonsmodellen og en modell som bare inneholder konstanten. Testen er oppgitt på Stata-utskriften:

$$\chi^2 = (-2) \times (\ln L_0 - \ln L_1) = (-2) \times (-4649,57 - -4587,57) = 124$$

Antall frihetsgrader er lik 1, fordi vi har inkludert en variabel. Vi ser at testen er statistisk signifikant. Det innebærer at en multinomisk regresjonsmodell med fagforeningsmedlemskap som avhengig variabel og kjønn som uavhengig gir en signifikant forbedring av loglikelihood enn dersom vi bare skulle basert oss på marginalfordelingen på den avhengige variabelen.

6.2 Eksempel 2: Fagforeningstilknytning, kjønn og alder

6.2.1 Regresjonsresultater og tolkning

Den bivariate eksempelet ovenfor viser sammenhengen mellom krysstabellanalyse og logistisk regresjon. Når vi i bivariat regresjonsanalyse har en dikotom uavhengig variabel er det mulig å estimere odds, oddsratioer og sannsynligheter som avviker relativt lite fra den opprinnelige fordelingen. Når vi trekker inn ytterligere uavhengige variabler, er det vanskeligere å studere sammenhengen mellom krysstabellanalyse og logistisk regresjon.

Nedenfor utvider vi eksempelet ovenfor med en ny forklaringsvariabel, nemlig alder. Resultatet av multinomisk logistisk regresjon er vist nedenfor.

```
. mlogit fagfor kjonn alder, b(0)

Iteration 0:  Log Likelihood = -4649.568
Iteration 1:  Log Likelihood =-4549.9611
Iteration 2:  Log Likelihood = -4548.337
Iteration 3:  Log Likelihood =-4548.3354

Multinomial regression                                Number of obs =   3209
                                                      chi2(8)       = 202.47
                                                      Prob > chi2   = 0.0000
Log Likelihood = -4548.3354                           Pseudo R2     = 0.0218

------------------------------------------------------------------------------
  fagfor |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Medlem L |
   kjonn |   .0521391   .0880787      0.592   0.554       -.120492    .2247702
   alder |   .0285957   .0041093      6.959   0.000       .0205416    .0366497
   _cons |   -1.14601   .1801964     -6.360   0.000      -1.499189   -.7928319
---------+--------------------------------------------------------------------
Medlem Y |
   kjonn |   1.105719   .1227231      9.010   0.000       .8651863    1.346252
   alder |   .0365768   .0055105      6.638   0.000       .0257765    .0473771
   _cons |  -3.051667   .2562164    -11.911   0.000      -3.553842   -2.549492
---------+--------------------------------------------------------------------
Medlem A |
   kjonn |  -.1136976   .1224627     -0.928   0.353      -.3537202    .1263249
   alder |   .0306986   .0055986      5.483   0.000       .0197256    .0416717
   _cons |  -2.221731     .25011     -8.883   0.000      -2.711937   -1.731524
---------+--------------------------------------------------------------------
Medlem f |
   kjonn |   .7775515   .1460953      5.322   0.000       .4912099    1.063893
   alder |   .0343935   .0066934      5.138   0.000       .0212747    .0475122
   _cons |  -3.293396   .3096175    -10.637   0.000      -3.900236   -2.686557
------------------------------------------------------------------------------
(Outcome fagfor==Ikke med is the comparison group)
```

Vi ser at kjønnskoeffisientene endrer seg relativt lite sammenliknet med den bivariate analysen. Det er fortsatt for YS og frittstående fagforeninger at vi finner signifikante kjønnsforskjeller. Det innebærer at oddsen for å være medlem av YS og frittstående fagforeninger snarere enn å ikke være organisert er høyere for kvinner enn for menn.

Alder gir et positivt, signifikant bidrag i samtlige likninger. Det innebærer at sjansen for å være organisert i en eller annen fagforening snarere enn å ikke være fagorganisert, synker med alderen.

6.2.2 Signifikanstester og «goodnes of fit»

Z-testen er som nevnt ikke signifikant for kjønnskoeffisienten i delmodellene som sammenlikner henholdsvis LO-organisasering og AF-organisering med det å ikke være organisert. Loglikelihood-testen for totalmodellen er signifikant. Det innebærer at vi kan forkaste en hypotese om at modellen ikke forbedrer log likelihood i forhold til en modell med bare en konstant. Kjikvadrattesten har to frihetsgrader fordi vi har inkludert to forklarinsvariabler.

Vi kan også teste om den trivariate modellen gir en forbedring av log likelihood, sammenliknet med den bivariate modellen. Vi tar forskjellen i log likelihood i de to modellene og ganger med –2:

$$\chi^2 = (-2) \times (\ln L_0 - \ln L_1) = (-2) \times (-4587{,}57 - -4548{,}34) = 78{,}46$$

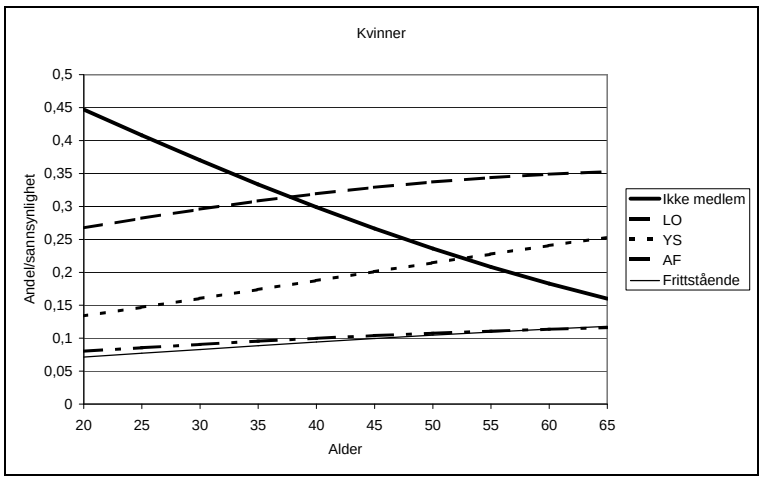
Antall frihetsgrader er lik 1. Forskjellen er følgelig signifikant. Den trivariate modellen forbedrer log likelihood sammenliknet med den bivariate modellen. Det er viktig å huske at forutsetningen for å kunne sammenlikne log likelihood i to ulike modeller er at den avhengige variabelen er den samme og at antall enheter er likt.

6.2.3 Beregning av sannsynligheter og presentasjon av resultater

Beregning av sannsynligheter er omfattende i multinomisk logistisk regresjon fordi vi først må beregne odds for alle de fire likningene for de gitte kombinasjonene av verdier på de uavhengige variablene. Operasjonen kan forenkles ved å bruke et regneark som f.eks. Excel, Lotus e.l. Nedenfor har vi vist resultatene av beregninger på bakgrunn av regresjonsresultatene for den trivatiate modellen. Kjønnsvariabelen er satt til verdien 1 (kvinner), mens aldersvariabelen varierer fra 20 til 65 år.

Regresjonskoeffisienter					
	Likning 1	Likning 2	Likning 3	Likning 4	
Kjønn	0,052	1,106	-0,114	0,778	
Alder	0,029	0,037	0,031	0,034	
Konstant	-1,146	-3,052	-2,222	-3,293	
Beregning av odds for kvinner					
Alder	Odds1	Odds2	Odds3	Odds4	
20	0,59809839	0,29939246	0,17978385	0,1596135	
25	0,69142541	0,36023453	0,20992604	0,18919066	
...	...	...	...	...	
60	1,90789397	1,3152148	0,62126348	0,62188506	
65	2,20560092	1,5824907	0,72542325	0,73712337	
Beregning av sannsynligheter for kvinner					
Alder	Ikke medlem	LO	YS	AF	Frittstående
20	0,44704961	0,26737965	0,13384328	0,0803723	0,07135515
25	0,40803392	0,28212502	0,14698791	0,08565695	0,07719621
...	...	...	...	...	...
60	0,18294053	0,34903113	0,24060609	0,11365427	0,11376798
65	0,15998366	0,35286011	0,25317266	0,11605587	0,1179277

En tabell med sannsynligheter er antakelig ikke spesielt givende. Det er bedre å illustrere sammenhengen mellom sannsynlighetene for å ha ulike verdier på den avhengige variabelen og en uavhengig variabel ved hjelp av en figur. Figuren nedenfor viser sammenhengen mellom fagforeningstilknytning og alder for kvinner.



Vi ser at sannsynligheten for å ikke være medlem av en fagforening synker etterhvert som alderen øker. Sannsynligheten for å være medlem i fagforening, enten LO, YS, AF eller frittstående øker derimot med alderen. På ethvert alderstrinn i figuren vil summen av sannsynlighetene for å falle i de fem kategoriene være 1.

7 Rangert (ordinal) logistisk regresjon

7.1 Problemet med variabler på ordinalnivå

En rekke sosiologisk interessante variabler er på ordinalnivå, dvs. at verdiene på variabelen kan rangeres, men at det er problematisk å snakke om avstanden mellom verdiene. Eksempler på slike variabler er alle slags former for rangeringer, hvor avstandene mellom verdiene ikke er eksakte:

- Holdninger til påstander: svært enig, nokså enig, litt enig, litt uenig, nokså uenig, svært uenig
- Tillit: stor tillit, litt tillit, liten tillit, svært liten tillit
- Karakterskala: svært godt, meget godt, godt, noe godt, lite godt
- Jobbvurderinger: skala fra 1 til 9, hvor 1 er svært lite tilfreds og 9 er svært tilfreds
- Sysselsetting: ingen, deltid, heltid
- Yrkesstatus: ufaglært arbeider, faglært arbeider, lavere funksjonær, midlere funksjonær, høyere funksjonær
- Alderskategorier: barn, ungdom, ung voksen, voksen, gammel
- Utdanning: grunnskole, videregående, høyskole, universitet

Det er ikke urimelig å hevde at svært mange sosiologiske variabler i realiteten befinner seg på ordinalnivå, dvs. at de uttrykker en rangering eller intensitet som ikke er presist kvantifiserbar. Østerberg peker for eksempel på at ordinalskalaen framstår:

> *«...som den mest egnede skala for statistisk behandling av sosiologiske variable. Det er en "intensiv" skala som måler "intensiteter"»* (1986:111).

Hellevik argumenterer for at både nominalvariabler, som bosted og yrke, og forholdstallsvariabler, som f.eks. inntekt, i mange tilfeller kan betraktes som variabler på ordinalnivå. Han peker på at:

> *«Hvilket målenivå det er rimelig å tillegge en variabel, avhengiger ikke bare av hvordan målingen er utført. Det vil også avhenge av hva slags teoretisk egenskap vi er ute etter å måle»* (Hellevik 1991:157-58)

Om en variabel formelt er på forholdstallsnivå, er det likevel ikke gitt at det i sosiologisk sammenheng alltid gir like stor mening å tillegge den et så høyt målenivå. Variabelen inntekt kan illustrere dette. Som et mål på økonomisk velferd, ville det være meningsløst å si at en økning i inntekten på kr. 10.000 betyr like mye for en minstepensjonist som for Kjell Inge Røkke. Som et mål på sosial status, ville det også gi lite mening i å spekulere på hvor mye endring i status en kan tilskrive en endring i inntekten på 10.000.

Samtidig er det naturligvis ikke gitt at en variabel skal behandles som en ordinalvariabel bare fordi den kan rangeres. Dette avhenger av problemstillingen og hensikten med analysen. Long (1997:115) nevner som eksempel på dette at selv om farger kan ordnes i henhold til det elektromagnetiske spektrum, er det mer enn tvilsomt om f.eks. variabelen fargevalg ved kjøp av bil bør behandles som en rangert variabel. Long trekker også fram problemer knyttet til variabler som uttrykker mer enn en dimensjon, f.eks. en holdningsskala med verdiene svært enig, enig, nøytral, uenig, svært uenig. I denne skalaen ligger det både retning på meningen (enig/uenig) og intensitet i holdningen. Dersom holdningens intensitet er av størst interesse, bør rangeringen være: svært (enig eller uenig), enig eller uenig og nøytral. Hvis det er uklart hvilken dimensjon en er ute etter bør variabelen behandles som om den var på nominalnivå.

Ordinalvariabler byr på problemer i statistisk analyse, spesielt når de opererer som avhengige variabler. Den beste analyseformen i problemstillinger har ofte vært krysstabellanalyse, med sine klare

begrensninger når det gjelder antall variabler som kan trekkes inn og hvor mange verdier disse variablene kan ha. Ofte har ordinalvariabler vært dikotomisert eller behandlet som om de var på forholdstallsnivå. I det første tilfellet mister en informasjon, i det andre tilfelle må en gjøre nokså sterke forutsetninger om avstandene mellom verdiene på variabelen.

La oss se litt nærmere på denne siste løsningsformen, fordi den gir en innfallsvinkel til prinsippet for ordinal logistisk regresjon. Hva innebærer det at man benytter en variabel på ordinalnivå som om den var på forholdstallsnivå?

Eksempel:
La oss ta som eksempel et tillitsspørsmål som ble stilt i 1996 om tillit til produsentenes miljømerking (Tufte og Lavik 1997): *I hvilken grad har du tillit til en at vare er miljøvennlig når produsentene garanterer for det?* Undersøkelsen ga følgende svarfordeling:

	Kode	*Prosent*
Meget stor tillit	5	5
Stor tillit	4	24
Både og	3	54
Liten tillit	2	13
Meget liten tillit	1	4
		100 (1008)

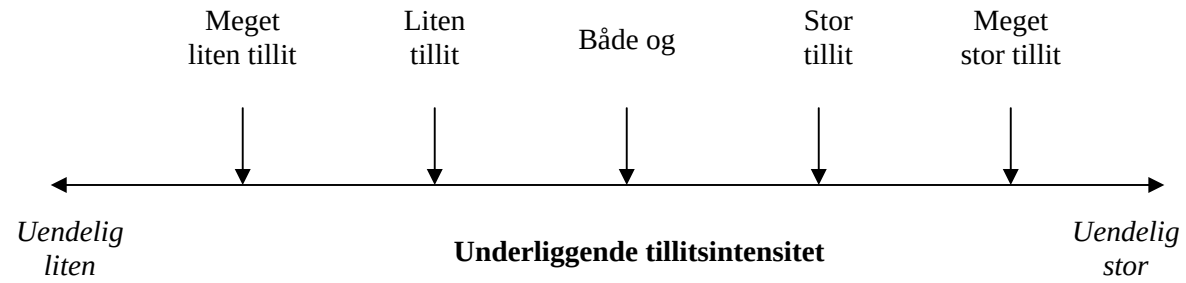
Verdiene på variabelen er kodet slik at kategorien «meget stor tillit» er gitt høyeste verdi, dvs. at de enhetene som havner i denne kategorien større tillit til produsentene enn de andre. For hvert nivå med lavere tillit synker den tilordnede tallverdien med 1. Denne tilordningen av tall markerer kun rangordningen mellom verdiene, ikke avstanden.

Dersom vi imidlertid behandler denne variabelen som om den var på forholdstallsnivå, vil vi også tillegge avstanden mellom verdiene vekt. Det innebærer at vi betrakter avstanden mellom «meget stor tillit» og «stor tillit» som like stor som avstanden mellom «liten tillit» og «meget liten tillit».

Men hva betyr egentlig avstandsbegrepet i forbindelse med en ordinalvariabel? Et tankeeksperiment kan være til hjelp her. La oss tenke oss at kategoriene på tillitsvariabelen ovenfor kun utgjør punkter på en underliggende, kontinuerlig tillitsvariabel, en variabel som vi ikke har målt og som antakelig heller ikke lar seg måle direkte. I praksis innebærer det at ordinalvariabelen måler en underliggende variabel på forholdstallsnivå.

Denne underliggende, ikke-målte variabelen er den «faktiske», kvantitative intensiteten (jf. sitatet fra Østerberg ovenfor) i folks tillit til produsentenes miljømerking. Denne kan antas å gå fra uendelig positiv verdi, dvs. uendelig stor tillit, til uendelig negativ verdi, dvs. uendelig liten tillit. Siden vi ikke har noen måte å måle denne tillitsintensiteten på (tillitsgenet er ennå ikke funnet), må vi heller ty til en ordinalskala med upresise kategorier som i høy grad er gjenstand for fortolkning av respondentene.

Hvis vi benytter variabelen ovenfor som om den var på forholdstallsnivå, innebærer det at vi forutsetter at kategoriene deler den underliggende tillitsintensiteten i like intervaller, slik det er illustrert i figuren nedenfor. Dersom denne forutsetningen er riktig, er det uproblematisk å behandle variabelen som om den var på forholdstallsnivå.



Forutsetningen om like intervaller er imidlertid nokså sterk i de fleste tilfeller. Sannsynligheten er svært stor for at avstandene varierer i størrelse. Dersom dette er tilfelle, er ikke betingelsene for ordinær regresjon tilfredsstilt.

En mulig løsning på dette problemet er å tilordne kategoriene verdier som gjør de underliggende avstandene mellom dem forskjellig. Tillitsvariabelen kunne f.eks. tilordnes verdier på følgende måte:

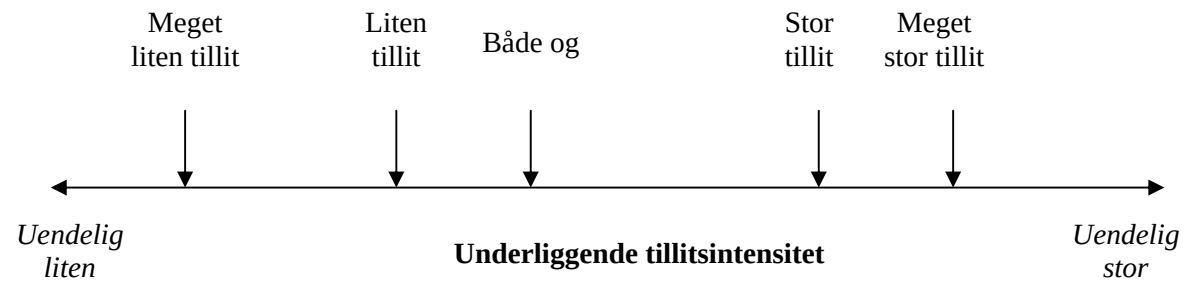
- Meget stor tillit - 3
- Stor tillit - 2
- Både og 0
- Liten tillit 2
- Meget liten tillit 3

Denne kodingen gir uttrykk for at sprangene fra å være nærmest likegyldig når det gjelder tillit til å ha enten stor tillit eller liten tillit er kvantitativt sett større enn sprangene fra å ha stor tillit til meget stor tillit eller fra å ha liten tillit til meget liten tillit. I prinsippet innebærer denne formen for koding at en gjør om ordinalvariabelen til en forholdstallsvariabel. Det er imidlertid sjelden vi har forutsetninger for å vurdere forholdet er mellom kategoriene på ordinalvariabelen og skalaen på den underliggende intensitetsvariabelen. Kodingen blir derfor svært skjønnsmessig, men det innebærer likevel ikke at denne framgangsmåten prinsipielt må avvises. Skjønn kan være bedre enn ren automatikk. Det er viktig å være klar over at det å beholde den opprinnelige kodingen også er et valg, nemlig at en forutsetter at avstandene mellom kategoriene på variabelen er like.

Dersom en mener at intervallene ikke er like, men samtidig ikke har grunnlag for å fastsette intervallene skjønnsmessig, er det en mulighet å benytte ordinal logistisk regresjon. Denne analyseformen forutsetter at den avhengige variabelen er på ordinalnivå, men estimerer avstanden mellom kategoriene på bakgrunn av observasjonene i datasettet og den spesifiserte regresjonsmodellen.

7.2 Prinsippet bak ordinal logistisk regresjon

Ordinal logistisk regresjon bygger på prinsippene bak logistisk regresjon med dikotom avhengig variabel. Verdiene på den avhengige variabelen må derfor gjøres om til et sett av dikotomier. Det gjøres ved å betrakte verdiene på ordinalvariabelen som *kuttpunkter* på en underliggende kontinuerlig variabel (Long 1977:116).



Kuttpunktet (verdien) *meget liten tillit* deler den underliggende variabelen slik at de som har tillitsintensitet lik eller mindre der kuttpunktet treffer linjen, har *meget liten tillit* til produsentene. De som har høyere tillitsintensitet, har *ikke* meget liten tillit, dvs. de kan ha alt fra liten tillit til meget stor tillit. Ved å sette skillet på denne måten har vi dikotomisert den opprinnelige variabelen.

Ordinal logistisk regresjon bygger imidlertid på at en foretar tilsvarende dikotomiseringer for alle verdiene på den avhengige variabelen. Den neste verdien eller kuttpunktet er *liten tillit*. Dette punktet skiller mellom på den ene siden dem som har liten eller meget liten tillit produsentene og på den andre siden dem som har fra både og til meget stor tillit. Setter vi skillet på denne måten får vi en ny dikotom variabel.

Ved å gjøre tilsvarende for samtlige verdier gjør vi om ordinalvariabelen til et sett av dummyvariabler som vist i tabellen nedenfor. For hver av dummyvariablene markerer verdien 1 høy tillitsintensitet og verdien 0 lav tillitsintensitet utfra kuttpunktet.

Verdier/kuttpunkter:	*Meget liten tillit*	*Liten tillit*	*Både og*	*Stor tillit*	*Meget stor tillit*
Ordinalvariabel	1	2	3	4	5
Dummyvariabler:					
Dummy 1	0	1	1	1	1
Dummy 2	0	0	1	1	1
Dummy 3	0	0	0	1	1
Dummy 4	0	0	0	0	1
Dummy 5	0	0	0	0	0

Den siste dummyvariabelen *Dummy 5* kan utelates fordi alle observasjoner her vil få verdien 0. De fem verdiene på ordinalvariabelen kan følgelig etter denne metoden gjøres om til 4 dummyvariabler, 1 variabel mindre enn det er verdier (Agresti 1990:321). Omformingen av den opprinnelig variabelen til et sett av dikotome variabler kalles for «collapsings of the response into binary outcomes» (Agresti 1996:212).

For hver av de dikotome variablene kan vi prinsipielt gjennomføre logistisk regresjon i henhold til framgangsmåten skissert i kapittel 2. For hver av dummyvariablene kan en beregne sannsynligheten for å ha verdien 1, basert på en regresjonsmodell hvor avhengig variabel er logaritmen av oddsen for å ha verdien 1. Vi får da fire regresjonsmodeller:

$$L_1 = \ln\left(\frac{P(\text{Dummy1}=1)}{1-P(\text{Dummy1}=1)}\right) = b_0 + b_1x_1 + ... + b_nx_n + e$$

$$L_2 = \ln\left(\frac{P(\text{Dummy2}=1)}{1-P(\text{Dummy2}=1)}\right) = b_0 + b_1x_1 + ... + b_nx_n + e$$

$$L_3 = \ln\left(\frac{P(\text{Dummy3}=1)}{1-P(\text{Dummy3}=1)}\right) = b_0 + b_1x_1 + ... + b_nx_n + e$$

$$L_4 = \ln\left(\frac{P(\text{Dummy4}=1)}{1-P(\text{Dummy4}=1)}\right) = b_0 + b_1x_1 + ... + b_nx_n + e$$

Aberet med framgangsmåten ovenfor er at vi får fire sett med koeffisienter, avhengig av hvilket kuttpunkt (dummyvariabel) vi analyserer. Dette gir en lite heldig måte å presentere resultatene på. Framgangsmåten er heller ikke tråd med de teoretiske forutsetninger, i og med at vi antar at ordinalvariabelen står i forhold til en underliggende kontimuerlig intensitetsvariabel. Vi har følgelig også en forestilling om at de enkelte uavhengige variabler virker på denne underliggende variabelen uttrykt ved ordinalvariabelen. Vi ønsker oss derfor ett sett med koeffisienter som uttrykk for denne påvirkningen, ikke fire ulike estimater for hver av de uavhengige variablene i modellen.

Ordinal logistisk regresjon går derfor et skritt videre ved å beregne alle disse fire likningene simultant under den beskrankning at hver variabel skal ha identiske koeffisientene i samtlige likninger. Noe upresist kan en si at forutsetningen for dette er at de uavhengige variablene har identiske effekter, målt i oddsratio, uansett hvor på skalaen en befinner seg.[20] Vi skal senere se på hva dette innebærer.

Det er på tide å gå litt inn på sannsynlighetene i ordinal logistisk regresjon. De er såkalt kumulative, dvs. at en hele tiden ser sannsynligheten for en ha en verdi eller høyere på ordinalvariabelen i forhold til sannsynligheten for å ha lavere verdier. Det innebærer at vi kan rangere sannsynlighetene:

sannsynligheten for å ha *meget stor* tillit til produsentene
P(dummy4=1)

er mindre enn

sannsynligheten for å ha *meget stor* eller *stor* tillit til produsentene
P(dummy3=1)

er mindre enn

sannsynligheten for å ha *meget stor, stor* eller *både og* tillit til produsentene
P(dummy2=1)

er mindre enn

sannsynligheten for å ha *meget stor*, *stor*, *både og* eller *liten tillit* til produsentene
P(dummy1=1)

er mindre enn

sannsynligheten for å ha *meget stor*, *stor*, *både og*, *liten* eller *meget liten* tillit til produsentene

som er lik 1.

I og med at sannsynlighetene rangeres kan også logaritmen av oddsene i likningene ovenfor rangeres. De utgjør følgelig kumultative logiter (Agresti 1990:321). På denne måten har vi spesifisert en relasjon mellom de fire likningene som kan danne utgangspunkt for beregning av et sett med koeffisienter. Det som er konstanter (b_0) i de fire likningene ovenfor, blir i ordinal logistisk regresjon betraktet som kuttpunkter. Disse brukes som utgangspunkt for å beregne sannsynligheter for å ha bestemt verdi på ordinalvariabelen gitt verdiene på de uavhengige variablene.

Beregningsmåten av ordinal logistisk regresjon er matematisk sett mer komplisert enn for dikotom logistisk regresjon, men følger samme prinsippet som denne. En bruker MLE for å finne estimater for effekter og kuttpunkter som gjør det observerte materialet mest sannsynlig gitt den modellen vi har spesifisert.

[20] For en mer utførlig utlegning av dette, se Agresti (1996:212).

Utskriften av analysen er nokså lik utskriften ved ordinær logistisk regresjon. Den eneste forskjellen er at en nå får estimater for kuttpunkter i stedet for en konstant. Den vesentligste forskjellen er hvordan koeffisienter skal fortolkes og sannsynligheter beregnes. Når det gjelder signifikanstester er det små forskjeller i forhold til binær logistisk regresjon. Når det gjelder goodness of fit er det større begrensninger. I prinsippet er det utelukkende mål knyttet til log likelihoodratio som lar seg beregne. Dette vil det bli gjort nærmere rede for i kapitlet om signifikanstesting og goodness of fit. I dette kapitlet vil vi konsentrere oss om fortolkning og beregninger av sannsynligheter, illustrert ved noen eksempler. Ved alle eksemplene har vi benyttet statistikkpakken Stata, fordi det ikke er mulig å foreta ordinal logistisk regresjon i SPSS.[21]

7.3 Eksempel 1: Lønn og kjønn

7.3.1 Innledende krysstabell – beregning av effekter

Eksempelet tar utgangspunkt i den samme lønnsvariabelen som tidligere, men i stedet for å inndele den i to verdier, som i kapittel 2, deles den nå inn i tre verdier: inntil 80 kroner timen, 80 til 115 kroner timen og 115 kroner eller mer i timen. Krysstabellen nedenfor viser den bivariate sammenhengen mellom lønn og kjønn.

```
. tabulate lonn_ord kjonn, col chi2 lrchi2

timelønn - | kjønn
tredelt    |      Mann     Kvinne |     Total
-----------+----------------------+----------
      < 80 |       218        532 |       750
           |     12.47      35.92 |     23.23
-----------+----------------------+----------
  80 - 115 |       885        797 |      1682
           |     50.63      53.81 |     52.09
-----------+----------------------+----------
     115 < |       645        152 |       797
           |     36.90      10.26 |     24.68
-----------+----------------------+----------
     Total |      1748       1481 |      3229
           |    100.00     100.00 |    100.00

          Pearson chi2(2) = 421.8266   Pr = 0.000
 likelihood-ratio chi2(2) = 446.2895   Pr = 0.000
```

Tabellen viser at prosentandelen med høy lønn er omtrent 27 prosentpoeng høyere for menn enn for kvinner. Når det gjelder kategorien middels lønn er forskjellen mellom menn og kvinner relativt liten. Andelen som har lav lønn er derimot klart størst blant kvinner. Sammenhengen i tabellen er monoton, dvs. at sammenhengen snur. Mens menn har den høyeste andelen høytlønnede, har kvinner den høyeste andelen lavtlønnede.

I stedet for å beregne prosentdifferanser kan vi bruke oddsratioer som mål på lønnsforskjellene mellom menn og kvinner. Vi kan sammenlikne oddsen for å befinne seg over et visst lønnsnivå for menn og kvinner ved å dele oddsen for kvinner med den tilsvarende oddsen for menn. Vi kan beregne følgende oddsratioer for tabellen ovenfor:

[21] Ved hjelp av loglineære modeller kan en også kjøre tilnærmet logistisk regresjon på ordinalvariabler i SPSS, men dette forutsetter kategoriske uavhengige variabler.

(1) Oddsratio for å ha høy eller middels lønn:

$$\theta = \frac{\text{Oddsen for kvinner}}{\text{Oddsen for menn}} = \frac{\frac{0,54+0,10}{0,36}}{\frac{0,51+0,37}{0,12}} = \frac{1,78}{7,33} = 0,24$$

Dette tallet innebærer at sjansen for å ha høy eller middels lønn (målt ved odds) for kvinner er 24% av den tilsvarende sjansen for menn.

(2) Oddsratio for å ha høy lønn:

$$\theta = \frac{\text{Oddsen for kvinner}}{\text{Oddsen for menn}} = \frac{\frac{0,10}{0,54+0,36}}{\frac{0,37}{0,51+0,12}} = \frac{0,11}{0,59} = 0,19$$

Fortolkningen av denne oddsratioen er at kvinner har en sjanse for å ha høy lønn som er 19 prosent av den tilsvarende sjansen for menn. En annen måte å kommentere denne oddsratioen på er å si at antall personer med høy lønn i forhold til antall personer med lav eller middels lønn er vesentlig lavere blant kvinner enn blant menn.

7.3.2 Regresjonsutskrift i STATA

Stata har en egen kommando for å gjennomføre ordinal logistisk regresjon som heter **.ologit**. Dersom vi gjennomfører ordinal logistisk regresjon mellom kjønn og den tredelte lønnsvariabelen får vi følgende resultat:

```
. ologit lonn_ord kjonn

Iteration 0:  Log Likelihood = -3306.934
Iteration 1:  Log Likelihood =-3090.9991
Iteration 2:  Log Likelihood =-3086.3697
Iteration 3:  Log Likelihood =-3086.3488

Ordered Logit Estimates                                 Number of obs =   3229
                                                        chi2(1)       = 441.17
                                                        Prob > chi2   = 0.0000
Log Likelihood = -3086.3488                             Pseudo R2     = 0.0667
------------------------------------------------------------------------------
lonn_ord |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   kjonn |  -1.490325   .0742623    -20.068   0.000      -1.635876   -1.344773
---------+--------------------------------------------------------------------
   _cut1 |   -2.03552   .0635271            (Ancillary parameters)
   _cut2 |    .565102   .0483433
------------------------------------------------------------------------------
```

Analysen gir en koeffisient for kjønnvariabelen. I tillegg får vi to kuttpunkter, ett kuttpunkt mindre enn antall verdier på den avhengige variabelen. Kuttpunktene og kjønnskoeffisienten kan nå benyttes til å beregne odds, oddsratioer og sannsynligheter.

Det første en kan merke seg er fortegnet på kjønnskoeffisienten. Koeffisienten er negativ, dvs. at kvinner har en større sjanse for å «havne« i den lave enden av skalaen enn menn, dvs. har at oddsen for å ha høy lønn er mindre for menn enn for kvinner. Dette er i tråd med analysen av krysstabellen ovenfor.

7.3.3 Beregning av odds og oddsratioer

Vi kan nå beregne blant annet beregne oddsen for å ha middels og høy lønn og oddsen for å ha høy lønn ved hjelp av følgende formel:

$$\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = e^{-(b_1x_1+\ldots+b_nx_n)-k}$$

Legg merke til at kuttpunktene fungere noe annerledes en konstant ved utregningen av oddsen. Mens konstanten legges til resultatet fra regresjonslikningen, trekkes kuttpunktet i fra.

(1) Odds for å ha høy eller middels høy lønn:

For å beregne dette må vi bruke det første kuttpunktet fordi dette skiller mellom lav lønn på den ene siden og middels/høy lønn på den andre. Det første kuttpunktet er estimert til –2,03. Beregnede odds blir da:

for menn: $$\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = e^{(0\times-1,49)--2,03} = e^{2,03} = 7,61$$

Fortolkning: Det er en sterk overvekt av personer med høy eller middels høy lønn blant menn.

for kvinner: $$\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = e^{(1\times-1,49)--2,03} = e^{0,54} = 1,72$$

Fortolkning: Det er også en overvekt av personer med høy eller middels høy lønn blant kvinner, men på langt nær så stor som hos menn. Det ser vi når vi beregner forholdet mellom oddsene, oddsratioen:

$$\theta = \frac{1,72}{7,61} = 0,23$$

Sjansen for å ha høy eller middels høy lønn for kvinner er bare 23% av den tilsvarende sjansen for menn. Når vi beregnet oddsratioen direkte ut fra krysstabellen fikk vi 0,24. Samsvaret mellom disse to beregningene er meget godt.

(2) Odds for å ha høy lønn:

For å beregne dette må vi bruke det andre kuttpunktet fordi dette skiller mellom lav/middels lønn på den ene siden og høy lønn på den andre. Det andre kuttpunktet er estimert til 0,57. Beregnede odds blir da:

for menn: $$\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = e^{(0\times -1,49)-0,57} = e^{-0,57} = 0,57$$

Fortolkning: Det er et mindretall blant mennene som har høy lønn. Andelen med høy lønn er 57 prosent av andelen som ikke har høy lønn.

for kvinner: $$\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = e^{(1\times -1,49)-0,57} = e^{-2,06} = 0,13$$

Fortolkning: Blant kvinner er det også et mindretall som har høy lønn og denne andelen er langt mindre enn blant menn. Igjen gjenspeiles dette i oddsratioen:

$$\theta = \frac{0,13}{0,57} = 0,23$$

Også her er kvinnenes sjanse for å ha høy lønn lik 23 prosent av mennenes sjanse. Når vi beregnet oddsratioen direkte fra krysstabellen fikk vi 0,19. Samsvaret er ikke så godt som for den første kategorien, men må likevel sies å være tilfredstillende.

Vi ser at regresjonsmodellen gir en rimelig bra tilnærming til odds og oddsratioer som vi beregnet ut fra krysstabellen ovenfor. Vi ser også at ordinal logistisk regresjon forutsetter, slik vi nevnte ovenfor, at effekten ,målt i log odds eller oddsratio, er konstant uansett for hvilke verdier på den avhengige variabelen vi beregner for. Effekten av kjønn, eller forskjellen mellom kvinner og menn målt i oddsratio, er hele tiden

$$\theta = e^{\beta} = e^{-1,49} = 0,23$$

dvs. antilogaritmen av kjønnskoeffisienten. Denne beregningsmåten er en raskere måte å komme fram til resultatene i regneeksemplene ovenfor.

7.3.4 Beregning av sannsynligheter

Som nevnt er ikke odds og oddsratioer enkle å forstå substansielt. Vi kan imidlertid innenfor ordinal logistisk regresjon beregne sannsynligheter på lik linje med det vi gjorde i kapittel 2. Formelen for dette er:

$$p = \frac{1}{1 + e^{(b_1x_1 + ... + b_nx_n) - k}}$$

Det er viktig å huske at det vi arbeider med i ordinal logistisk regresjon er kumulative sannsynligheter. For å beregne sannsynligheten for å havne i laveste lønnskategori bruker vi kuttpunkt 1. For å beregne sannsynligheten for å havne i laveste eller midterste lønnskategori bruker vi kuttpunkt 2. Sannsynligheten for å være i laveste, midterste eller høyeste lønnsgruppe er selvfølgelig lik 1 (det finnes ingen andre muligheter).

Beregningen av sannsynligheter skjer i to trinn:

(1) Først beregnes resultatet av uttrykket:

$$(b_1x_1 + ... + b_nx_n) - k$$

hvor k betegner det aktuelle kuttpunktet.

(2) Deretter settes dette resultatet inn i formelen:

$$p = \frac{1}{1 + e^{(b_1x_1 + ... + b_nx_n) - k}}$$

Dermed er den kumulative sannsynligheten beregnet

I vårt eksempel innebærer det følgende beregning av sannsynligheter:

	Mann	Kvinne
Sannsynlighet for lav lønn (kuttpunkt = -2,04)	$p = \frac{1}{1 + e^{-1,49 \times 0 - (-2,04)}} = 0,12$	$p = \frac{1}{1 + e^{-1,49 \times 1 - (-2,04)}} = 0,37$
Sannsynlighet for lav eller middels lønn (kuttpunkt = 0,57)	$p = \frac{1}{1 + e^{-1,49 \times 0 - 0,57}} = 0,64$	$p = \frac{1}{1 + e^{-1,49 \times 1 - 0,57}} = 0,89$

De kumulative sannsynlighetene må deretter regnes om til sannsynligheter:

	Menn	*Kvinner*
Sannsynlighet for *lav lønn* = ingen beregning nødvendig	0,12	0,37
Sannsynlighet for *middels lønn* = sannsynlighet for lav og middels lønn minus sannsynlighet for lav lønn	0,64 – 0,12 = 0,52	0,89 – 0,37 = 0,52
Sannsynlighet for *høy lønn* = 1 minus sannsynlighet for lav og middels lønn	1 – 0,64 = 0,36	1 – 0,89 = 0,11

Vi ser at de beregnede sannsynlighetene stemmer svært godt overens med frekvensene i tabellen. Det vil imidlertid ikke nødvendigvis alltid være tilfelle.

7.3.5 Signifikanstester

Kjønnskoeffisienten har i følge regresjonskoeffisienten en asymptotisk standardfeil på 0,07. Dersom vi deler koeffisienten på standardfeilen får vi en z-verdi på –20,01. Dete innebærer at kjønnskoeffisienten er signifikant forskjellig fra null og vi får støtte for en hypotese om det er en sammenheng mellom kjønn og lønnsnivå.

Som tidligere nevnt forutsetter denne testen et stort utvalg. Alternativt kan vi derfor teste forskjellen i log-likelihood mellom den aktuelle modellen og en modell som bare inkluderer konstanten (jf. avsnitt 4.1). Denne testen er også oppgitt på STATA-utskriften. Forskjellen er:

$$\chi^2 = (-2) \times (\ln L_0 - \ln L_1) = (-2) \times (-3306{,}93 - -3086{,}35) = 441{,}16$$

Denne er signifikant med en frihetsgrad (fordi vi har inkludert en variabel). Det innebærer at modellen med kjønn som uavhengig variabel gir en sterkt signifikant bedre log-likelihood enn en modell med bare en konstant.

Vi ser også at Stata oppgir pseudo-R^2 for ordinal logistisk regresjon. Som nevnt er denne vanskelig å fortolke og bør benyttes med forsiktighet, fortrinnsvis til å sammenlikne hva som skjer når en inkluderer flere uavhengige variabler i en modell.

7.4 Eksempel 2: Lønn, kjønn og utdanning

7.4.1 Utskrift og tolkning av koeffisienter

Vi utvider nå det bivariate eksempelet ovenfor ved å trekke inn ytterligere en uavhengig variabel, nemlig utdanning. Antakelsen bak modellen er at sannsynligheten for å havne i ulike lønnskategorier henger sammen forskjeller i kjønn og utdanningslengde. Nedenfor ser vi utskriften fra Stata:

```
. ologit lonn_ord kjonn utdaar

Iteration 0:  Log Likelihood = -3213.787
Iteration 1:  Log Likelihood =-2732.3212
Iteration 2:  Log Likelihood =-2714.5398
Iteration 3:  Log Likelihood =-2714.2557
Iteration 4:  Log Likelihood =-2714.2556

Ordered Logit Estimates                                 Number of obs =   3152
                                                        chi2(2)       = 999.06
                                                        Prob > chi2   = 0.0000
Log Likelihood = -2714.2556                             Pseudo R2     = 0.1554

------------------------------------------------------------------------------
lonn_ord |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   kjonn |  -1.489244   .0772989    -19.266   0.000      -1.640747   -1.337741
  utdaar |   .3482134    .015877     21.932   0.000        .317095    .3793318
---------+--------------------------------------------------------------------
   _cut1 |  -1.270273    .073142             (Ancillary parameters)
   _cut2 |   1.743937   .0755401
------------------------------------------------------------------------------
```

Kjønnskoeffisienten endrer seg ikke når vi inkluderer utdanning. Det innebærer at det er liten sammenheng mellom de to uavhengige variablene. Fortolkningen er at kvinner har en større sjanse for å ha lav verdi og mindre sjanse for å ha høy verdi på den avhengige variabelen enn menn. Oddsen for å ha høy lønn er mindre for kvinner enn for menn, mens oddsen for å ha lav lønn er større.

Utdanningskoeffisienten er positiv. Det innebærer at jo lengre utdanning respondentene har, jo mer sannsynlig er det at de befinner seg i det øvre lønnsnivået. Antall respondenter med høy lønn i forhold til de andre øker med utdanning. Det samme gjør forholdstallet mellom antall respondenter med høy og middels høy lønn og de med lav lønn.

Begge koeffisientene er sterkt signifikante (z-test). Vi får med andre støtte for hypotesene om at lønnsnivå henger sammen med både kjønn og utdanning. Alternativt kan vi støtte oss på kjikvadratesten av forskjellen i log likelihood mellom modellen ovenfor og en modell med bare en konstant. Vi ser at denne forskjellen er signifikant med to frihetsgrader (vi har inkludert to variabler i modellen).

Det er også mulig å teste forskjellen i log-likelihood mellom den bivariate modellen vi analyserte i forrige avsnitt og den trivariate modellen ovenfor. Denne testen forutsetter at antallet observasjoner er likt i de to modellene. Dette er ikke tilfelle i vårt eksempel, noe som antakelig skyldes at en del respondenter ikke har oppgitt utdanningslengde. Vi kan derfor derfor ikke gjøre denne testen med mindre vi tar ut de manglende observasjonene også i den bivariate modellen (slik vi gjorde i eksempelet i kapittel 2). Stat kan teste forskjellen i log likelihood mellom ulike modeller ved hjelp av kommandoen **.lrtest**.[22]

7.4.2 Beregning av sannsynligheter og grafisk presentasjon

Sammenhenger mellom avhengig variabel og uavhengige variabler på forholdstallsnivå kan det være interessant å presentere grafisk. En må da beregne sannsynligheter for ulike verdier på den aktuelle uavhengige variabelen, gitt en bestemt kombinasjon av verdier på de andre uavhengige variablene. I vårt eksempel kan det være interessant å illustrere sammenhengen mellom utdanning og lønnsnivå for henholdsvis menn og kvinner. Formlene for å beregne sannsynligheter er de samme som i forrige

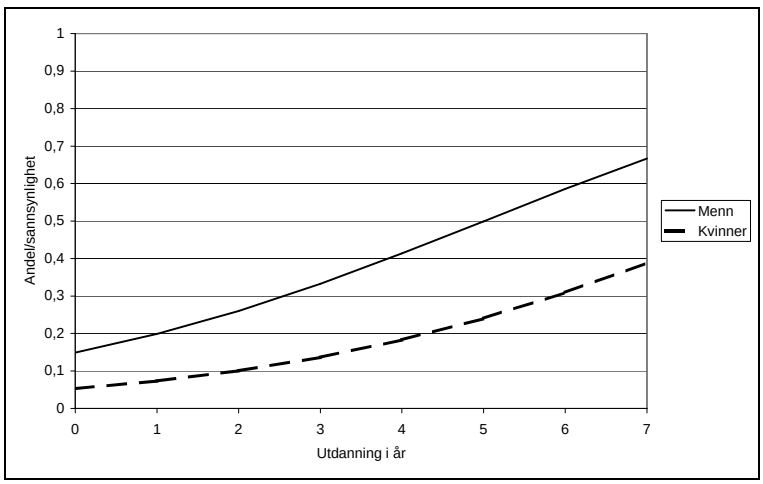
[22] Et eksempel på dette er vist i kapittel 4.

eksempel, men nå må både kjønns,- og utdanningsvariabelen tas i betraktning. Dette blir fort mange beregninger, slik at det er mest hensiktsmessig å foreta dem i et regneark, for eksempel Excel. Nedenfor ser vi et oppsett for beregning av sannsynligheter basert på modellen ovenfor.

Variabel	**Estimat**					
Kjønn	-1,1489244					
Utdanning	0,3482134					
Kutt 1	-1,270273					
Kutt 2	1,743937					
Sannsynligheter						
Kjønn	Menn			Kvinner		
Lønnsnivå	Lav	Middels	Høy	Lav	Middels	Høy
Utdanning 0	0,21921052	0,63197592	0,14881355	0,46970002	0,4777924	0,05250758
1	0,16541372	0,63609069	0,19849559	0,38471992	0,54249322	0,07278686
2	0,1227438	0,6175527	0,2597035	0,30623608	0,59369279	0,10007112
3	0,08989536	0,57813648	0,33196815	0,23758079	0,62633732	0,13608188
4	0,06518455	0,52169572	0,41311972	0,18031686	0,63725878	0,18242436
5	0,04691619	0,45380131	0,4992825	0,13442174	0,62541678	0,24016148
6	0,03358372	0,38092837	0,58548791	0,09880005	0,59193992	0,30926003
7	0,02394483	0,30929749	0,66675769	0,07183469	0,54007998	0,38808533

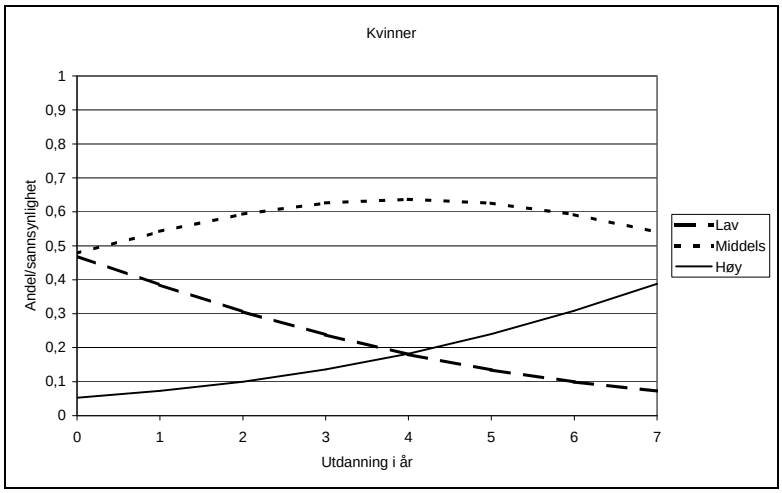
Legg merke til at summen av sannsynligheten for å ha lav, middels og høy lønn for en gitt kombinasjon av kjønn og utdanning (kovariat) alltid blir 1. For menn med 1 års utdanning er for eksempel summen av sannsynlighetene: 0,22+0,63+0,15=1.

Det er mange måter å presentere sannsynlighetene ovenfor på. Dersom vi ønsker å sammenlikne menn og kvinner i ett eneste diagram, kan det enkleste være å velge en bestemt verdi på den avhengige variabelen og illustrere sannsynlighetskurven for denne. Dersom vi for eksempel velger verdien *høy lønn*, kan vi lage følgende diagram som viser sammenhengen for henholdsvis menn og kvinner:



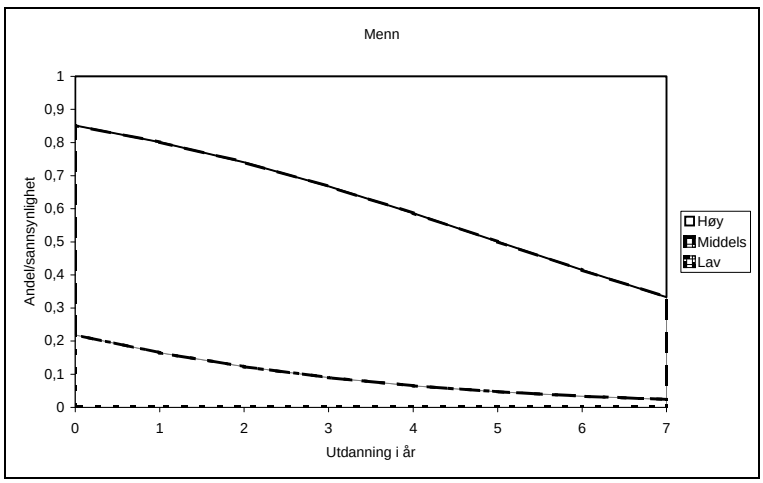
Figuren viser at sannsynligheten for å ha høy lønn stiger med utdanning både for menn og kvinner. Menn har en høyere sannsynlighet, men stigningen for kvinner er sterkere etter hvert. Figuren gir et ryddig bilde, men en mister detaljinformasjon om de andre verdiene på den avhengige variabelen.

Alternativt kan vi tegne inn alle verdiene, men det er da vanskelig å få med sammenhengen for både menn og kvinner i samme diagram. Nedenfor ser vi utviklingen i sannsynlighet for å ha henholdsvis lav, middels og høy lønn for kvinner:



Vi ser at den predikerte andelen som har lav lønn synker relativt raskt, mens andelen med høy lønn stiger. Sannsynligheten for å tilhøre den midterste gruppen stiger fram til 4 års utdanning for deretter å synke.

En annen måte å illustrere sammenhengen på kan være å bruke kumulative diagrammer, hvor linjene i diagrammet legger seg oppå hverandre, slik det er vist nedenfor. Dette kan være en fin måte å illustrere «sammensetningen» eller den betingede fordelingen på den avhengige variabelen ved ulike verdier på den uavhengige variabelen.



Vi ser at andelen med lav lønn blant menn er forholdsvis liten, selv ved lav utdannelse og at den synker ytterligere når utdanningen øker. Ved utdanning = 0 er det gruppen med middels lønn som dominerer, men utdanning lik 6 år eller høyere, dominerer gruppen med høy lønn.

8 Oppsummering

Notatet heter en intuitiv innføring i logistisk regresjon. Dette er antakelig en tittel som vil provosere fordi notatet tross alt inneholder mye formler og beregninger. Det som forsvarer tittelen er likevel at hovedprinsippet bak logistisk regresjon er forsøkt forklart mest mulig intuitivt. Den matematikken som er inkludert, er hovedsakelig motivert ut fra behovet for å gjøre en del egne beregninger når en gjør logistisk regresjon. Det er ikke noe problem å få et statistikkprogram til å gjøre tilsvarende beregninger. Stata har f.eks. funksjoner som kan brukes til å gjøre de samme beregningene og ikke minst lage plot av de beregnede sannsynligheter. Å få til gode plot er imidlertid vanskelig når antallet uavhengige variabler øker, noe som henger sammen med at plot i Stata baserer seg på observasjoner i materialet. Fordelen med å bruke et regneark, som Lotus eller Excel, er at en kan framstille selve regresjonslikningen, uavhengig av hvor mange observasjoner som tilfredsstiller de kriterier vi setter.

Tabellen nedenfor viser en oppsummering av forholdet mellom de tre former for logistisk regresjon som er gjennomgått i dette notatet.

Type logistisk regresjon	*Målenivå på avhengig variabel*	*Flere koeffisienter pr. variabel*	*Omgjøring av avhengig variabel*
Binær	Dikotom	Nei	Nei
Multinomisk	Nominal	Ja	Odds mellom enkeltverdier og referanseverdi
Ordinal	Ordinal	Nei	Odds mellom kumulativ sannsynligheter

Hovedprinsippet er at avhengige variabler med mer enn to verdier må gjøre om til flere dikotome avhengige variabler. Det medfører at man i multinomisk logistisk regresjon får flere koeffisienter for hver variabel, mens man i ordinal logistisk regresjon får en koeffisient.

Logistisk regresjon byr på store fordeler når det gjelder analyse av «kvalitative» variabler. Samtidig er det viktig å ha i mente at også denne analyseteknikken bygger på forutsetninger. Den viktigste forutsetningen er formen på den avhengige variabelen – S-kurven. Dersom det ikke er meningsfylt at sannsynlighetene/andelene følger denne kurven, må andre modeller benyttes.

Det største problemet med logistisk regresjon er formidling. Resultatene er ikke intuitivt lette å forstå og stiller store krav til presentasjon. Vi har viet en stor del av dette notatet til denne siden av formidlingsarbeidet. Formidling byr på ekstraarbeid for forskeren, men en vil som regel oppleve at ulike framstillingsteknikker også byr på økt innsikt. I det hele tatt krever logistisk regresjon en mer kritisk og aktiv holdning fra forskeren på flere måter. I tillegg til utfordringer når det gjelder presentasjon gir ikke logistisk regresjon enkle mål på goodnes of fit slik OLS gjør. På mange måter er dette bare en fordel siden det øker bevisstheten omkring modellspesifiseringen. Det er viktig å teste ulike modeller opp mot hverandre, men då må det gjøres ved hjelp av likelihoodtesten og ikke ved blind maksimering av R^2.

Litteratur

Achen, Christopher H. (1982): *Interpreting and using regression.* Sage Publications, Newbury Park.

Agresti, Alan (1990): *Categorical Data Analysis.* John Wiley & Sons, New York.

Agresti, Alan (1996): *An Introduction to Categorical Data Analysis.* John Wiley & Sons, New York.

Aldrich, John H. & Forrest D. Nelson (1984): *Linear Probability, Logit, and Probit Models.* Sage, Beverly Hills.

Freeman, Daniel H. (1987): *Applied Categorical Data Analysis.* Marcel Dekker, New York.

Greene, William H. (1993): *Econometric Analysis.* Macmillan, New York.

Galtung, Johan (1970): *Theory and Mehods of Social Research.* Universitetsforlaget, Oslo.

Gujarati, Damodar N. (1988): *Basic Econometrics.* McGraw-Hill, New York.

Hagquist, Curt & Magnus Stenbeck (1998): «Goodness of Fit in Regression Analysis – R^2 and G^2 reconsidered» i *Quality and Quantity. International Journal of Methodology*, vol. 32, no 3, august 1998.

Hamilton, Lawrence (1998): *Statistics with Stata 5.* Duxbury Press, New Hampshire.

Hellevik, Ottar (1991): *Forskningsmetode i sosiologi og statsvitenskap.* Universitetsforlaget, Oslo.

Hernes, Gudmund (1976): «En intuitiv innføring i multivariat analyse» i Stein Ugelvik Larsen (red.): *Problemer i samfunnsvitenskapelig metode*, Universitetsforlaget, Oslo.

Kmenta, Jan (1986): *Elements of Econometrics.* Macmillan, New York.

Knoke, David & Petter J. Burke (1980): *Log-linear models.* Sage, Newbury Park.

Lewis Beck, M.S. (1980): *Applied Regression. An Introduction.* Sage, Beverly Hills.

Lillestøl, Jostein (1991): *Sannsynlighetsregning og statistikk.* Bedriftsøkonomen Forlag, Oslo.

Long, Scott J. (1997): *Regression Models for Categorical and Limited Dependent Variables*. Advanced quantitative techniques in the social sciences, vol 7, California.

Maddala, G. S. (1983): *Limited dependent and qualitative variables in econometrics.* Cambridge University Press, New York.

Skog, Ole Jørgen (1998): *Å forklare sosiale fenomener. En regresjonsbasert tilnærming.* Ad Notam Gyldendal, Oslo.

SPSS (1993): *Advanced Statistics 6.1.* SPSS Inc., Chicago.

StataCorp. (1999): *Stata Reference Manual. Release 6.0.* Volume 2 H-O, Stata Press, Texas.

Sørensen, Rune (1989): «Logitmodellen: Analyse av diskret avhengig variabel.» i *Tidsskrift for samfunnsforskning.* Årgang 30, s. 61-86.

Tufte, Per Arne og Randi Lavik (1997): *Helse- og miljøinformasjon. Forbrukernes behov for informasjon om skadelige stoffer i produkter.* Rapport nr. 4-1997. Statens institutt for forbruksforskning, Lysaker.

Østerberg, Dag (1986): *Fortolkende sosiologi.* Universitetsforlaget, Oslo.