2006/54

Notater

Statistics Norway

Statistisk sentralbyrå

Johan Heldal

Logistisk regresjon - kurskompendium i byråskolens kurs SM507

Seksjon for metoder og standarder

Innhold

1. Hva er regresjon?

Først: hva er enkel lineær regresjon? Mange av dere vil være familiære med begrepet eller i det minste har dere vært borti det i studiene. Begrepet har sitt opphav hos Francis Galton som i 1885 publiserte en artikkel han kalte "Regression Toward Mediocracy in Hereditary Stature". Galton studerte der sammenhengen mellom fedres og sønners høyde og fant ut at høye fedre hadde en tendens til i gjennomsnitt å få høye sønner, og lave fedre lave sønner, men sønnene hadde en tendens til ikke å være like høye/lave som fedrene. Ordet regresjon er blitt hengende ved den metoden Galton brukte, selv om det strengt tatt er en misvisende term. Regresjon er en felles betegnelse på metoder der en studerer sammenhengen mellom statistiske variable med sikte på å forklare variasjonen i en variabel som vi kan kalle Y ved hjelp av en eller flere andre variable som vi kan kalle $x_1, x_2, \ldots, x_k$.

I klassisk lineær regresjon er Y en kontinuerlig (skala) variabel som f.eks. en persons høyde, vekt, inntekt, blodtrykk osv. eller en bedrifts omsetning, profitt, mengde utslipp (av en eller annen art) . Y kalles "responsvariabelen", "den avhengige variable", "den endogene variable" e.l., litt avhengig av terminologien til det fagområdet metoden anvendes innenfor. $x_1, \ldots, x_k$ kalles tilsvarende for "uavhengig variabel", "stimuli", "forklaringsvariabel", "exogen variabel" e.l. Disse er ofte også kontinuerlige, men behøver ikke være det. Kjønn er en mye brukt forklaringsvariabel som ikke er kontinuerlig. Et eksempel med en Y og en x variabel er følgende tabell over teaterbesøk og billettinntekter.

Tabell 1 Billettinntekter og besøk for 22 teatre i 2002.[1]

Navn	Nr.	Besøk	Inntekt	Navn	Nr.	Besøk	Inntekt
Rolle	i	x_i	Y_i		i	x_i	Y_i
Agder Teater	1	74958	9383	Hålogaland Teater	12	38127	4453
Beaivvàs Sàmi Teahter	2	14922	572	Nationaltheatret	13	167287	31975
Black Box Teater	3	15430	2614	Nordland Teater	14	35780	1035
Brageteatret	4	15532	347	Nord-Trøndelag Teater	15	12572	742
Carte Blanche Danseteater	5	7930	956	Oslo Nye Teater	16	172046	22601
Den Nationale Scene	6	90154	9683	Riksteatret	17	90846	10818
Den Norske Opera	7	155934	36699	Rogaland Teater	18	67614	8535
Det Norske Teatret	8	188890	22296	Sogn og Fjordane Teater	19	32165	1257
Haugesund Teater	9	35839	1561	Teater Ibsen	20	26940	2616
Hedmark Teater	10	22478	955	Teatret Vårt	21	38092	3179
Hordaland Teater	11	21304	1401	Trøndelag Teater	22	95306	13270

I en enkel lineær regresjonsmodell postulerer vi en lineær sammenheng med feil mellom x og Y variablene, dvs.

$$Y = \alpha + \beta x + e \qquad (1.1)$$

der forventningen (gjennomsnittet) til e er null: $E(e) = 0$, og variansen er σ^2, $V(e) = \sigma^2$. Vi kan også skrive dette som

$$E(Y \mid x) = \alpha + \beta x, \quad V(Y \mid x) = \sigma^2 . \qquad (1.2)$$

(forventet bilettinntekt). $E(Y \mid x)$ kalles den *betingede forventningen* gitt verdien til x. $V(Y \mid x)$ er den *betingede variansen*. Symbolet "$\mid x$" leses som "gitt x". $E(Y \mid x)$ (i 1000 kroner) hvis x (antall besøk) øker med en enhet:

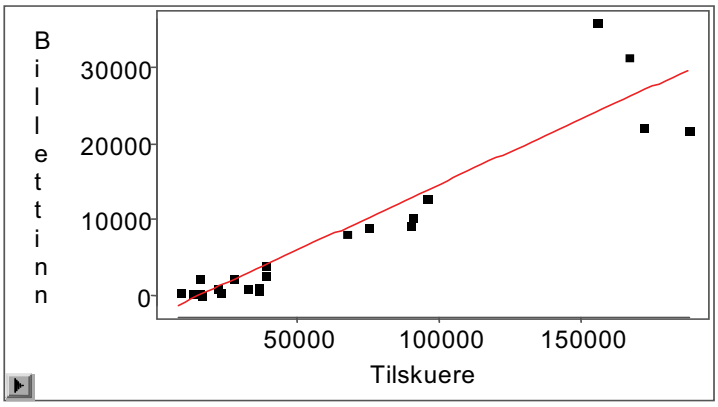
[1] Kilde: Statistisk årbok 2003 tabell 279.

$$E(Y \mid x+1) - E(Y \mid x) = (\alpha + \beta(x+1)) - (\alpha + \beta x) = \beta \qquad (1.3)$$

Tolkningen av parameteren α er som verdien av $E(Y, x)$ hvis $x = 0$:

$$E(Y \mid x = 0) = \alpha + \beta \cdot 0 = \alpha. \qquad (1.4)$$

På grunnlag av observasjonene i tabell 1 kan α, β og σ^2 estimeres som $\hat{\alpha} = -2534.37, \hat{\beta} = 0.1715$ og $\sigma^2 = 15532332.1$ $(\sigma = 3941.1)$ Tabellen kan plottes i et spredningsplott sammen med en linje som representer $\alpha + \beta x$ og hvis stigning er representert ved β. Et slikt spredningsplott produsert ved SAS INSIGHT er er vist i figur 1:



Hva er gjennomsnittlig billettpris beregnet på denne måten?

Som indikert i begynnelsen, kan man ha flere forklaringsvariable i en regresjon. Det kalles da multivariabel regresjon og modellen blir da formulert som

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e \qquad (1.5)$$

der k er antall forklaringsvariable som er med i modellen.

2. Hva er logistisk regresjon?

Logistisk regresjon er aktuell når responsvariabelen Y er en *kategorivariabel*, en variabel hvis verdimengde er et endelig antall mulige kategorier. En kategorivariabel kan være *nominal* eller *ordinal*. Hos en ordinal kategorivariabel har kategoriene en naturlig ordningsrekkefølge, mens hos nominale variable har de det ikke. Eksempler på nominale og ordinale kategorivariable finnes i tabell 2.

Tabell2 Ulike typer kategoriske variable.

Variabel	**Kategorier**	**Type**
Kjønn	Mann, kvinne	Likegyldig
Sysselsettingsstatus	I arbeid, arbeidssøkende, utenfor arbeidsstyrken	Ordinal
Partipreferanse ved valg	Partiene/listene som stiller	Vanligvis nominal, men partiene kan også ordnes, f. eks langs høyre-venstre akse. Ordningen mellom partiene ikke opplagt.
Inntekt kategorisert	F.eks. ≤50000, 50-100000, 100-200000, 200-300000, >300000	Ordinal
Sykdomssymptom	F.eks. Ingen, milde, moderate, alvorlige.	Ordinal
Holdningsspørsmål	F eks. Likert skala, helt uenig --- helt enig.	Ordinal
Næringhovedområde	Næringshovedområder i NACE	Nominal
Yrkeshovedgrupper		Nominal

Variable kan ha noen kategorier som er ordinale og noen som faller utenfor ordningen, f.eks. ”Uoppgitt”. Alle ordinale variable kan i logistisk regresjon behandles som nominale ved å se bort fra ordningen. Det er gjerne enklest, men man kaster egentlig bort informasjon på den måten. Det finnes spesialmodeller som prøver å ivareta ulike ordinale strukturer i responsvariabelen. PROC LOGISTIC, som er den SAS-prosedyren som vil få mest oppmerksomhet i dette kompendiet, antar uten videre at responsvariabelen er ordinal og anvender en av disse spesialmodellene, kalt "parallell lines regression" eller også "proportional odds model". Det er nå også mulig å kjøre PROC LOGISTIC med nominal responsvariabel med flere enn to kategorier, såkalt generalisert logistisk regresjon. For dette formål er det også mulig å bruke PROC GENMOD. Denne vil imidlertid ikke bli behandle i dette kompendiet.

For responsvariable som bare har to svarkategorier, *dikotome* variable, vil spørsmålet om ordinal eller nominal variabel være irrelevant. I dette kurset vil vi for det meste beskjeftige oss med dikotome responsvariable, men komme inn på ”proportional odds” modellen til slutt. For dikotome variable vil det være hensiktsmessig å kode svarkategoriene med $Y = 0$ og $Y = 1$. Eksempler på slike dikotome responsvariable kan være:

Tabell 3: Noen dikotome responsvariable

Variabel	Verdier
Internett hjemme	0 = nei, 1 = ja
Arbeidsufør	0 = nei, 1 = ja
Hjertesykdom	0 = nei, 1 = ja
Svarstatus i undersøkelse	0 = frafall, 1 = respondent

Anta at sannsynligheten for å observere $Y = 1$ vil avhenge av verdien til en (eller flere) forklaringsvariable x. La $\pi(x)$ betegne denne sannsynligheten, dvs. $\pi(x) = P(Y = 1 \mid x)$. Da vil

$$E(Y \mid x) = 1 \cdot P(Y = 1 \mid x) + 0 \cdot P(Y = 0 \mid x) = P(Y = 1 \mid x) = \pi(x) . \qquad (2.1)$$

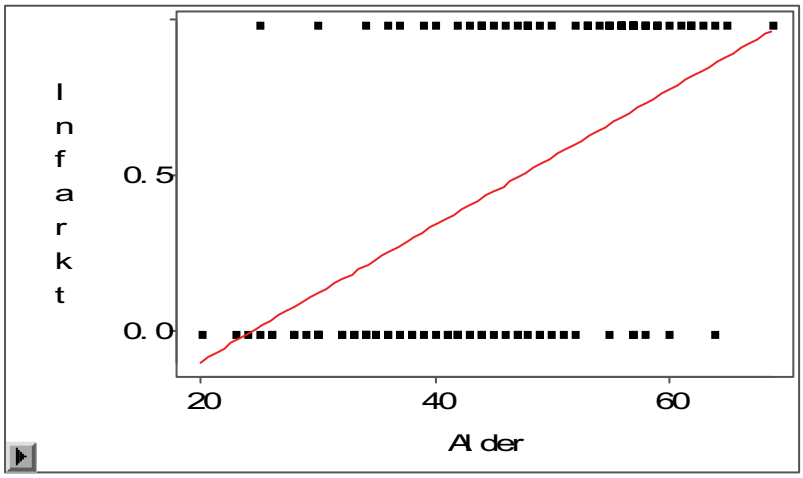
$\pi(x)$ vil på en eller annen måte avhenge av forklaringsvariablene x. Vi kunne tenke oss at $\pi(x) = \alpha + \beta x$ akkurat som i (1.2) for vanlig lineær regresjon. Det er ett spesielt problem med det. Siden $\pi(x)$ er en sannsynlighet, må den være et tall mellom 0 og 1. Uttrykket $\alpha + \beta x$ er ikke nødvendigvis et tall mellom 0 og 1. Hvis $\beta > 0$ og $x < -\alpha/\beta$ eller $\beta < 0$ og $x > -\alpha/\beta$ blir $\alpha + \beta x < 0$. Tilsvarende, hvis $\beta > 0$ og $x > (1 - \alpha)/\beta$ eller $\beta < 0$ og $x < (1 - \alpha)/\beta$ blir $\alpha + \beta x > 1$. Dette kan gå bra hvis x-verdier som gir $\alpha + \beta x > 1$ eller < 0 ikke er aktuelle, men det viser seg svært ofte at slike verdier er svært aktuelle.

***Eksempel* 1**: Tabell 4 viser et eksempel med $n = 100$ personer der personenes alder er x-variabel og infarkt er Y-variabel.

Tabell 4. 100 deltagere i en studie etter indisier på hjerteinfarkt. Y er 1 indikerer hjerteinfarkt, 0 ikke. AGRP er en aldersgruppering i 8 grupper. Tabellen er hentet fra boken *Applied Logistic Regression*, Hosmer & Lemeshow (1989).

	x	Y	#		x	Y	#
AGRP	Alder	Infarkt	Antall	AGRP	Alder	Infarkt	Antall
1	20	0	1	5	46	0	1
1	23	0	1	5	46	1	1
1	24	0	1	5	47	0	2
1	25	0	1	5	47	1	1
1	25	1	1	5	48	0	1
1	26	0	2	5	48	1	2
1	28	0	2	5	49	0	2
1	29	0	1	5	49	1	1
2	30	0	5	6	50	0	1
2	30	1	1	6	50	1	1
2	32	0	2	6	51	0	1
2	33	0	2	6	52	0	1
2	34	0	4	6	52	1	1
2	34	1	1	6	53	1	2
3	35	0	2	6	54	1	1
3	36	0	2	7	55	0	1
3	36	1	1	7	55	1	2
3	37	0	2	7	56	1	3
3	37	1	1	7	57	0	2
3	38	0	2	7	57	1	4
3	39	0	1	7	58	0	1
3	39	1	1	7	58	1	2
4	40	0	1	7	59	1	2
4	40	1	1	8	60	0	1
4	41	0	2	8	60	1	1
4	42	0	3	8	61	1	1
4	42	1	1	8	62	1	2
4	43	0	2	8	63	1	1
4	43	1	1	8	64	0	1
4	44	0	2	8	64	1	1
4	44	1	2	8	65	1	1
5	45	0	1	8	69	1	1
5	45	1	1				

I et vanlig spredningsplott vil Y-verdiene nå ligge langs to linjer, $Y = 0$ og $Y = 1$. som figur 2 viser.



Enkel lineær regresjon av samme type som i avsnitt 1 gir en regresjonslinje som passerer 0 i $x = -\hat{\alpha}/\hat{\beta} = -(-0.5380)/0.0218 = 24.68$ og 1 i $x = (1-\hat{\alpha})/\hat{\beta} = (1-(-0.538))/0.0218 = 70.55$.

Det er mange funksjonsformer som kan brukes for $\pi(x)$ for å sikre at den holder seg mellom 0 og 1. Logistisk regresjon svarer til at en benytter funksjonsformen

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{2.2}$$

eller hvis en har multippel logistisk regresjon:

$$\pi(\mathbf{x}) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}. \tag{2.3}$$

(2.3) kan omskrives til formen

$$\text{logit}\,\pi(\mathbf{x}) \equiv \log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k \tag{2.4}$$

der log vanligvis står for naturlig logaritme. Dette kalles *logit*-transformasjonen til $\pi(\mathbf{x})$. Fet **x** betyr $(x_1, \ldots, x_k)$, vektoren av verdiene til alle x-variablene. Forholdet $\pi(\mathbf{x})/(1-\pi(\mathbf{x}))$ kalles *oddsen* for verdien $Y = 1$ når forklaringsvariablene tar verdien **x**. Begrepene odds og log odds (logit) er helt sentrale for å tolke parametrene i en logistisk regresjonsmodell. For tenk deg at du gir en av x-variablene, si x_2, et tillegg på 1 slik at den øker med en enhet, så vil

$$\begin{aligned}
&\log\frac{\pi(\mathbf{x}+1_2)}{1-\pi(\mathbf{x}+1_2)} - \log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \\
&= (\alpha+\beta_1 x_1+\beta_2(x_2+1)+\beta_3 x_3+\cdots+\beta_k x_k) - (\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_k x_k) \qquad (2.5)\\
&= \beta_2
\end{aligned}$$

Her betyr $\mathbf{x} + 1_2$ at $\mathbf{x}$ har fått et tillegg på 1 i variabel 2 men er ellers uforandret. Men ifølge regnereglene for logaritmer er

$$\log\left(\frac{\pi(\mathbf{x}+1_2)}{1-\pi(\mathbf{x}+1_2)} \Big/ \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \log\frac{\pi(\mathbf{x}+1_2)}{1-\pi(\mathbf{x}+1_2)} - \log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \beta_2 . \qquad (2.6)$$

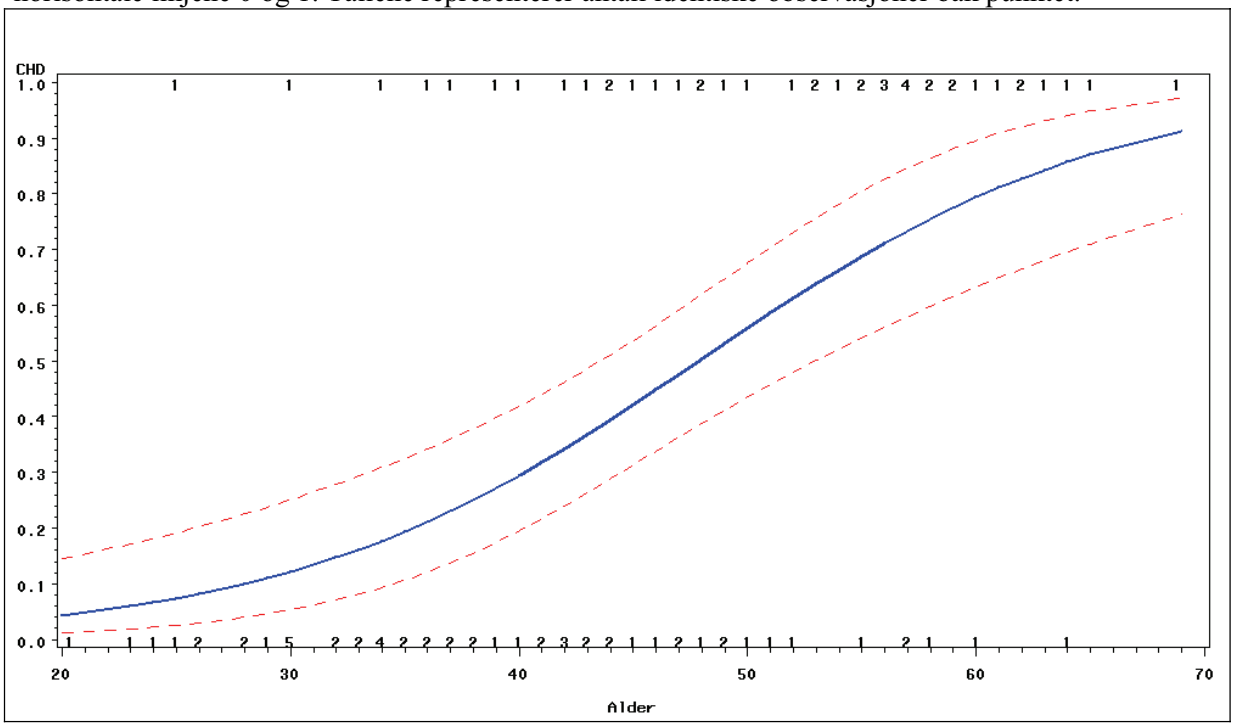
β_2 blir altså lik logaritmen til forholdet mellom to oddser, log til en *oddsrate*. e^{β_2} blir oddsraten selv. Merk at mens $\pi(\mathbf{x})$ kan variere mellom 0 og 1 vil odds($\mathbf{x}$) = $\pi(\mathbf{x})/(1-\pi(\mathbf{x}))$ variere fra 0 til ∞ og logit $\pi(\mathbf{x}) = \log(\pi(\mathbf{x})/(1-\pi(\mathbf{x})))$ vil variere fra $-\infty$ til ∞. Hvis $\pi(\mathbf{x}) = ½$ blir odds($\mathbf{x}$) = 1 og logit $\pi(\mathbf{x}) = 0$.

Logit funksjonen kalles en *link* funksjon. Alternative link funksjoner er `probit` transformasjonen og `identity` link. `identity` link er den elementære $\pi(\mathbf{x}) = \alpha + \beta x$ som ble demonstrert i eksempel 1 ovenfor. `probit` link bruker $\pi(x) = \Phi(\alpha + \beta x)$ der $\Phi(z)$ er den kumulative fordelingsfunksjonen til normalfordelingen, den som tabelleres i standard normalfordelingstabeller. Link funksjonen blir da den "inverse" funksjonen Φ^{-1}: $\Phi^{-1}(\pi(\mathbf{x})) = \alpha + \boldsymbol{\beta x}$. Ennå en annen link funksjon er *komplementær log-log*: $\log(-\log(1-\pi(\mathbf{x}))) = \alpha + \boldsymbol{\beta}\mathbf{x}$. PROC LOGISTIC kan håndtere både probit og komplementær log-log. I tillegg har LOGISTIC nå fått en link som kalles `glogit` for å gjøre det mulig å kjøre generalisert logistisk regresjon, dvs. logistisk regresjon hvor responsvariabelen har mer enn to nominale kategorier. Denne er omtalt i "The Logistic Procedure Update" lagt ut på Q:\METODEKURS\SM07\DOCS.

Eksempel fortsetter: Kjøring av den logistiske regresjonsmodellen for hjerteinfarktdataene i PROC LOGISTIC og grafen i figur 3 med PROC GRAPH gir blant annet følgende resultat:

```
Logit(infarkt) = -5.3095 + 0.1109 alder
```

Figur 3. Plott av $\pi(x)$ definert ved (2.7) med 95% konfidensbånd. De originale data ligger på de horisontale linjene 0 og 1. Tallene representerer antall identiske observasjoner bak punktet.



SAS program med PROC LOGISTIC for estimering av modellen og for plotting samt utskrift er gjengitt i avsnitt 5.1.1 og 5.1.2. Visse sammenhenger mellom størrelsene i utskriften er beskrevet.

Den estimerte $\pi(x)$ får formen

$$\hat{\pi}(x) = \frac{e^{-5.3095+0.1109x}}{1+e^{-5.3095+0.1109x}}. \qquad (2.7)$$

I (2.5) og (2.6) ovenfor viste vi at når forklaringsvariabelen x eller en variabel i den multiple forklaringsvariabelen $\mathbf{x}$ øker med en enhet, vil logitfunksjonen, log oddsen, øke med β eller med den parameteren β_j som hører til den komponenten x_j av $\mathbf{x}$ som øker.

Et naturlig spørsmål er da: Hvor mye øker $\pi(x)$? Svaret får vi ved å derivere $\pi(x)$ med hensyn på x, og det gir

$$\pi'(x) = \frac{d}{dx}\pi(x) = \beta\pi(x)(1-\pi(x)). \qquad (2.8)$$

I det multiple tilfellet må vi differensiere med hensyn på den enkelte komponent, som gir

$$\frac{\partial}{\partial x_j}\pi(\mathbf{x}) = \beta_j\pi(\mathbf{x})(1-\pi(\mathbf{x})). \qquad (2.9)$$

Siden kurven for $\pi(\mathbf{x})$ ikke er en rett linje blir økningen avhengig av x (eller $\mathbf{x}$). Økningen er minst når $\pi(\mathbf{x})$ er svært liten eller svært stor. Den blir størst når $\pi(\mathbf{x}) = ½$. da er $\pi(\mathbf{x})(1- \pi(\mathbf{x})) = ¼$. Å utlede (2.8) og (2.9) er en passende oppgave for dem som behersker derivasjon.

Det er vanlig ved logistisk regresjon å *gruppere* x-variablene før analysen. Dette tegner ofte et klarere bilde på en enkel måte og er ofte nyttig ved bestemmelse av modell for regresjonen. Slik gruppering gjør det mulig å fremstille data i en frekvenstabell og å formulere den logistiske regresjonsmodellen som en *log-lineær* modell. En logistisk regresjonsmodell basert på bare kategoriske/grupperte forklaringsvariable er ekvivalent med en log-lineær modell. Dette er ofte til stor hjelp når vi har flere forklaringsvariable og ønsker å bestemme en regresjonsmodell som passer til data. Vi vil alltid ønske å kunne finne en regresjonsmodell som er så enkel som mulig men som likevel gir en adekvat beskrivelse av de observerte data. Variabelen AGRP i tabell 4 definerer en gruppering av aldersvariabelen i eksempel 1. Ved hjelp av den kan vi redusere tabell 4 (en individfil) til tabell 5:

Tabell 5: Frekvenstabell for aldersgruppe x infarkt og skårer.

Alders gruppe	*n*	Infarkt		Andel ja	Mulige skårer		
		Nei	Ja		Midt-punkt	Gj.-snitt	Median
20-29	10	9	1	0.10	25.0	25.9	26.0
30-34	15	13	2	0.13	32.5	32.5	32.5
35-39	12	9	3	0.25	37.5	37.4	37.5
40-44	15	10	5	0.33	42.5	42.8	42.5
45-49	13	7	6	0.46	47.5	47.6	47.5
50-54	8	3	5	0.63	52.5	52.4	52.5
55-59	17	4	13	0.76	57.5	57.4	57.5
60-69	10	2	8	0.80	65.0	63.5	63.0
Total	100	57	43	0.43			

For at sammenhengen med log-lineære modeller skal holde og denne sammenhengen skal kunne utnyttes fullt ut, er det av betydning at antall grupper og grensene mellom dem er bestemt på forhånd (fast), uavhengig av hvor mange observasjoner en får og ikke etter at en har tittet i data før en bestemmer inndelingen. Er en x-variabel i utgangspunktet kontinuerlig kan det være fristende å

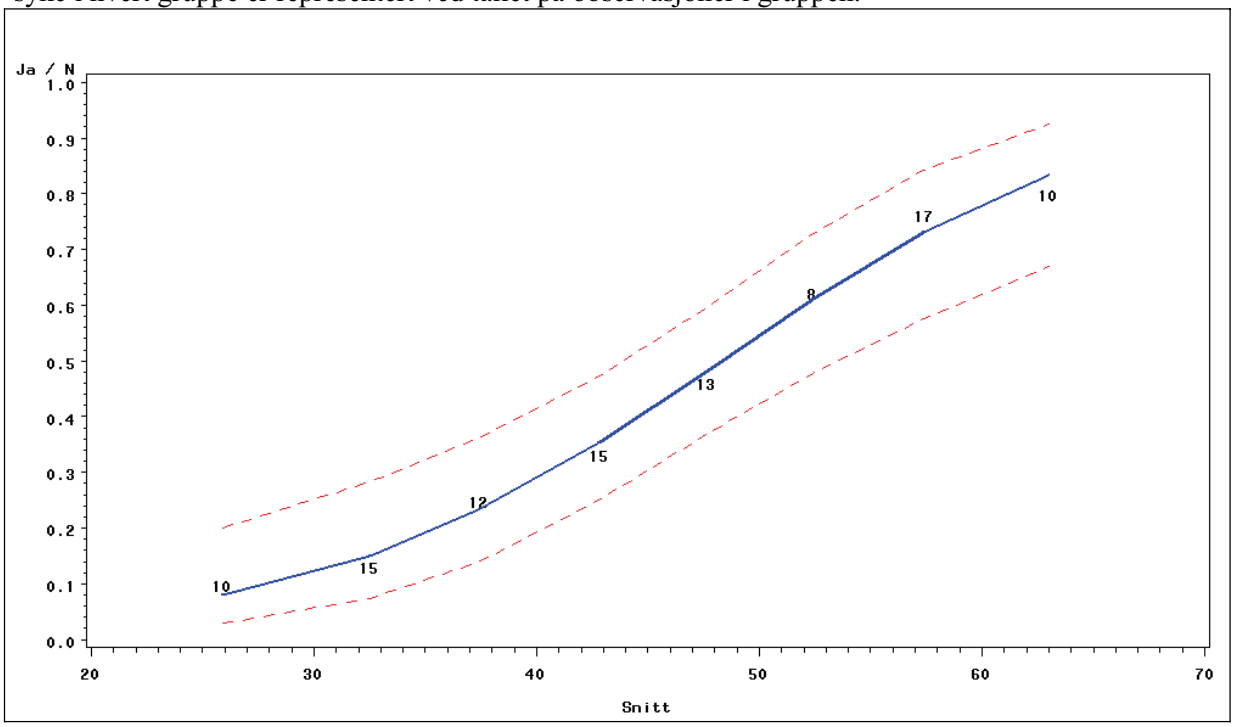
gruppere like x-verdier i data sammen. En må da huske på at like x-verdier gjerne opptrer fordi målenøyaktigheten på x-variabelen er med et begrenset antall siffer. Antall grupper ville bli tilfeldig og avhengig av data, og jo flere observasjoner en har jo flere grupper ville en måtte regne med å få.

Likevel kan det være ønskelig noen ganger å dele inn gruppene etter at data er observert. En kan være interessert i å gruppere med like mange i hver gruppe, for eksempel etter kvartiler eller desiler i x-verdiene i data. Hosmer og Lemeshow anviser en slik automatisk metode som kan brukes approksimativt.

Når en grupperer data må en velge skårer for de ulike gruppene til erstatning for de opprinnelige x-verdiene. Det er vanlig å bruke gruppemidtpunktene, men gjennomsnitt eller median for x-ene i hver gruppe kan også benyttes. Ved beregning av gjennomsnitt og median for alder i tabell 5 er det lagt til 0.5 år til hver persons alder i tabell 4 for å reflektere at alder er mer enn bare antall fylte år. Et plott av andelene med infarkt i hver gruppe basert på gjennomsnittsalder er vist i figur 4.

```
Logit(infarkt) = -5.2869 + 0.1094 alder
```

Figur 4. Plott av $\pi(x)$ med 95% konfidensbånd for grupperte data. Punktene for de originale andelene syke i hvert gruppe er representert ved tallet på observasjoner i gruppen.



Estimatene for de logistiske regresjonsparametrene blir ikke vesentlig forskjellige fra dem en fikk ved ugrupperte data. SAS program med PROC LOGISTIC for estimering av modellen og for plotting samt utskrift er gjengitt i avsnitt 12.1.3.

Når alle **x**-variablene er kategoriske, eventuelt etter å ha blitt gruppert, kan en også kjøre en logistisk regresjon ved å bruke programmer for log-lineære modeller. En bruker da heller PROC GENMOD. Dette er nyttig når en skal sammenligne modeller der forskjellige forklaringsvariable inngår. En får da mulighet til å teste modellens tilpasning tildata ved å sammenligne med en "mettet" modell. Dette blir først et viktig poeng når en har flere forklaringsvariable i **x**.

3. Logistisk regresjon og prosentvise endringer

3.1. 2 x 2 tabeller

I det aller enkleste tilfellet er forklaringsvariabelen x i en logistisk regresjonsmodell selv en dikotom variabel. I slike tilfeller kan vi redusere datasettet til en enkel 2 x 2 tabell. I elementære statistikkurs på universitetet lærer de fleste først å analysere slike tabeller ved å se på "prosentvise forskjeller" i responsvariabelen mellom de to nivåene av forklaringsvariabelen. Slike 2 x 2 tabeller er spesielt egnet til å illustrere forskjellene mellom slike prosentanalyser på den ene siden og logistiske/log-lineære analyser på den andre. I visse akademiske miljøer har det forekommet en aktiv debatt om hva slags type analyser som er det best egnede analyseverktøyet. Jeg tør likevel si at av dem som kjenner logistisk regresjon og log-lineære modeller er det få som ikke velger disse.

***Eksempel 1*:** De fleste eksemplene i dette kompendiet er tatt fra lastebilundersøkelsen med data fra 1. kvartal 2003. Vi kan begynne med en 2 x 2 tabell for hvorvidt en bil har svart i undersøkelsen (`deltok`) og om den kjører egenkjøring eller leiekjøring (`kj_art2`). Spørsmålet er om det er noen sammenheng mellom hva slags type kjøring det er og tilbøyeligheten for å svare i undersøkelsen.

```
Table of deltok by kj_art2

Y = deltok(Deltatt i undersøkelsen(0=Nei,1=Ja))
X = kj_art2(0:egen, 1:leie)
```

Frequency Row Pct Col Pct	X = 0	X = 1	Total
Y = 0	135 36.49 13.96	235 *q0*=**63.51** 10.12	370
Y = 1	832 28.51 *p0*=**86.04**	2086 *q1*=**71.49** *p1*=**89.88**	2918
Total	967	2321	3288

Mange vil, ut fra det de har lært, studere slike sammenhenger ved å se på prosentvise sammenligninger, dvs. sammenligne de estimerte sannsynlighetene for å svare (andelene som svarer) ved hhv leiekjøring (p_1) og egenkjøring (p_0) og se på differensene $p_1 - p_0 = 89{,}88 - 86{,}04 = 3{,}84\%$. Så kan man teste om denne observerte prosentdifferensen er statistisk signifikant. (Det gjøres ikke her.)

Nå kunne man tenke seg å snu problemstillingen og sammenligne andelene som kjører leietransport hhv blant dem som har svart (q_1) og dem som ikke har det (q_0) og se på differensene $q_1 - q_0 = 71{,}49 - 63{,}51 = 7{,}98\%$ og teste signifikansen av denne.

Nå kan en spørre seg: Hvis en med disse differensene ønsker å gi et mål på *sammenhengen* mellom de to variablene, bør ikke målene for sammenheng være et samme uansett fra hvilken synsvinkel vi ser den?

La oss prøve å analysere sammenhengen på grunnlag av odds og oddsforhold i stedet. Fra den første synsvinkelen vil først se på odds for å delta blant dem som kjører leiekjøring, $Odds(Y=1 \mid X=1) = p_1/(1-p_1) = 2086/235 = 8{,}88$ mot tilsvarende odds for dem som kjører egenkjøring, $Odds(Y=1 \mid X=0) = p_0/(1-p_0) = 832/135 = 6{,}16$. Så kan vi sammenligne disse oddsene ved å ta forholdet mellom dem:

$$\frac{Odds(Y=1\,|\,X=1)}{Odds(Y=1\,|\,X=0)}=\frac{p_1/(1-p_1)}{p_0/(1-p_0)}=\frac{2086/235}{832/135}=\frac{2086\cdot 135}{832\cdot 235}=1{,}4403$$

Hvis vi nå ser dette fra motsatt synsvinkel, vil vi sammenligne odds for å kjøre leietransport for biler som har svart $Odds(X=1\mid Y=1) = 2086/832 = 2{,}50$ mot tilsvarende odds for dem som ikke har svart, $Odds(X=1\mid Y=0) = 235/135 = 1{,}74$. Tar vi så forholdet mellom disse oddsene får vi

$$\frac{Odds(X=1\,|\,Y=1)}{Odds(X=1\,|\,Y=0)}=\frac{q_1/(1-q_1)}{q_0/(1-q_0)}=\frac{2086/832}{235/135}=\frac{2086\cdot 135}{235\cdot 832}=1{,}4403,$$

altså akkurat det samme. Dette oddsforholdet kaller vi også *kryssproduktet* i tabellen, fordi det etter fjerning av de brudne brøkene kan beskrives som produktet av tallene på (hoved)diagonalen dividert med produktet av tallene utenfor diagonalen. Jeg vil i det følgende bruke begrepene kryssprodukt og oddsforhold som synonyme.

Nå kunne vi tenke oss at vi hadde stratifisert utvalget annerledes slik at vi fikk dobbelt så mange med egenkjøring. Vi ville da kunne ha observert en tabell som så slik ut:

Y = deltok(Deltatt i undersøkelsen(0=Nei,1=Ja))
X = kj_art2(0:egen, 1:leie)

Frequency Row Pct Col Pct	***X*** = 0	***X*** = 1	Total
Y = 0	270 53.46 13.96	235 *q0*=**46.54** 10.12	505
Y = 1	1664 44.37 *p0*=**86.04**	2086 *q1*=**55.63** *p1*=**89.88**	3750
Total	1934	2321	4255

En ser at differensen $p_1 - p_0 = 3{,}84\%$ fortsatt, mens differensen $q_1 - q_0 = 55{,}63 - 46{,}54 = 9{,}09\%$, altså endret. Hva skjer med oddsratene? Fra den første synsvinkelen får vi

$$\frac{Odds(Y=1\,|\,X=1)}{Odds(Y=1\,|\,X=0)}=\frac{p_1/(1-p_1)}{p_0/(1-p_0)}=\frac{2086/235}{1664/270}=\frac{2086\cdot 270}{1664\cdot 235}=1{,}4403$$

som er det samme som før. Fra den andre synsvinkelen får vi

$$\frac{Odds(X=1\,|\,Y=1)}{Odds(X=1\,|\,Y=0)}=\frac{q_1/(1-q_1)}{q_0/(1-q_0)}=\frac{2086/1664}{235/270}=\frac{2086\cdot 270}{235\cdot 1664}=1{,}4403$$

som også blir uforandret.

Det vi effektivt har gjort er å multiplisere den første kolonnen i tabellen med 2. Det forandret prosentdifferensen i den ene retningen, men ikke i den andre. Det forandret *ikke* oddsforholdet, som fra før var det samme i begge retninger. Vi kunne ha multiplisert begge radene og/eller begge kolonnene med hvert sine tall. Det ville ha forandret på marginalfordelingene i tabellen og/eller de prosentvise endringene. Det ville ikke endret oddsforholdet.

Endringer i en tabell som kan beskrives ved at en eller flere rader eller kolonner multipliseres uniformt med faste tall kaller vi *marginale endringer* i tabellen. Det forholdet at marginale endringer ikke påvirker kryssproduktet gjør at vi - i denne forstand - kan se kryssproduktet som et mål for sammenheng mellom variablene i tabellen som er uavhengig av deres marginale fordelinger. Når det så også er uavhengig av synsvinkel, kan vi med rette si at oddsforholdet er et mye mer genuint mål på sammenheng mellom variablene i tabellene enn prosentvise differenser. I tabeller med flere rader og kolonner kan vi lage kryssprodukt for hvilken som helst 2x2 subtabell. Det gjelder da det samme, nemlig at disse kryssproduktene ikke påvirkes av marginale endringer.

Hva har dette med logistisk regresjon å gjøre? Tenk nå at type kjøring (eie eller leie) er en forklaringsvariabel, kall den x, og at vi koder $x = 0$ hvis egenkjøring og $x = 1$ hvis leiekjøring. La $\pi(0)$ og $\pi(1)$ være sannsynligheten for å svare i undersøkelsen i de to tilfellene. Vi kan sette opp den logistiske regresjonsmodellen

$$\text{logit}\,\pi(x) = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x \qquad (3.1)$$

Setter vi $x = 0$, estimerer vi $\pi(0)$ med $p_0 = 832/967 = 0{,}8604$. Setter vi dette videre inn i (3.1), får vi

$$\text{logit}\,p_0 \equiv \log\frac{832/967}{1-832/967} \equiv \log\frac{832}{135} \equiv \log 6{,}1630 \equiv 1{,}8146 = \hat{\alpha} + \hat{\beta}\cdot 0 \equiv \hat{\alpha}\,.$$

ˆ over symbolet viser at det er et estimat.
Setter vi $x = 1$, estimerer vi $\pi(1)$ med $p_1 = 2086/2321 = 0{,}8988$. Setter vi dette videre inn i (3.1), får vi

$$\text{logit}\,p_1 \equiv \log\frac{2086/2321}{1-2086/2321} \equiv \log\frac{2086}{235} \equiv \log 8{,}8766 \equiv 2{,}1834 = \hat{\alpha} + \hat{\beta}\cdot 1 \equiv \hat{\alpha} + \hat{\beta}$$

Vi estimerer β med

$$\hat{\beta} = (\hat{\alpha} + \hat{\beta}) - \hat{\alpha} = \text{logit}\,p(1) - \text{logit}\,p(0) = \log\frac{2086}{235} - \log\frac{832}{135} = \log\frac{2086/235}{832/135}$$
$$= \log\frac{2086\cdot 135}{832\cdot 235} = \log 1{,}4403 = 0{,}3649$$

Med andre ord: $\hat{\beta}$ er logaritmen til kryssproduktet i tabellen. (Vi regner med naturlige logaritmer). Dette viser det nære forholdet mellom logistisk regresjon og oddsforhold i en 2 x 2 tabell.

Det er også forholdsvis lett å regne ut et (asymptotisk) estimat for standardfeilen til $\hat{\beta}$:

$$SE(\hat{\beta}) = \sqrt{\frac{1}{2086} + \frac{1}{135} + \frac{1}{832} + \frac{1}{235}} = 0{,}1155$$

Wald Chi-Square som skrives ut av PROC LOGISTIC kan beregnes som

$$W = (\hat{\beta}/SE(\hat{\beta}))^2 = (0{,}3649/0.1155)^2 = 9{,}976$$

95% konfidensintervall for β kan beregnes som

$$\hat{\beta} \pm 1{,}96 SE(\hat{\beta}) = 0{,}3649 \pm 1{,}96\cdot 0{,}1155 = 0{,}3649 \pm 0{,}2264\,.$$

Det beregnede 95% konfidensintervallet blir dermed (0.1384, 0.5913). Det tilsvarende konfidensintervallet for selve oddsen blir $(e^{0,1384}, e^{0,5913}) = (1.1485, 1,8062)$.

4. Kategoriske forklaringsvariable med flere kategorier

4.1. Teori

I slike tilfeller må vi representere kategoriene med "dummy" variable. En variabel med f.eks. 6 kategorier må representeres ved en vektor $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$ der $x_i = 1$ for en av komponentene og 0 for de øvrige. Den logistiske regresjonsmodellen kan skrives

$$\text{logit}\,\pi(\mathbf{x}) = \log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \alpha + \beta_1 x_1 + \cdots + \beta_6 x_6 \qquad (4.1)$$

Nå er problemet at $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 1$, slik at hver av x-ene kan uttrykkes som 1 minus summen av de øvrige. For eksempel blir $x_6 = 1 - x_1 - x_2 - x_3 - x_4 - x_5$. Setter vi dette inn i (4.1) får vi

$$\begin{aligned}\text{logit}\,\pi(\mathbf{x}) = \log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} &= \alpha + \beta_1 x_1 + \cdots + \beta_5 x_5 + \beta_6(1 - x_1 - \cdots - x_5) \\ &= \alpha + \beta_6 + (\beta_1 - \beta_6)x_1 + (\beta_2 - \beta_6)x_2 + \cdots + (\beta_5 - \beta_6)x_5\end{aligned} \qquad (4.2)$$

Det er 6 regresjonsparametre men bare 5 selvstendige forklaringsvariable. Dette er et eksempel på det som kalles et *kolinearitetsproblem* og fører til at det ikke uten videre blir mulig å bestemme $\beta_1, \ldots, \beta_6$ entydig i forhold til hverandre. For å løse problemet må man innføre *restriksjoner* på parametrene. Det er her uendelig mange muligheter, men to som er de mest vanlige:

1. *Sum-til-null* parametrisering: $\beta_1 + \beta_2 + \cdots + \beta_6 = 0$, eller
2. *Referansekategori*parametrisering: Velg en kategori som referansekategori og sett den tilhørende β lik 0.

Når en kjører logistisk regresjon med f.eks. PROC LOGISTIC er det viktig å være oppmerksom på hvilken parametrisering programmet har brukt når en skal tolke resultatene. Hvis en ikke ber om noe spesielt vil PROC LOGISTIC bruke sum-til-null parametrisering som standard. I PROC LOGISTIC kalles dette *effect coding*. En kan imidlertid be om andre parametriseringer. Hvis en velger referanse-kategoriparametrisering og ikke spesifiserer hvilken kategori en vil ha som referansekategori, vil PROC LOGISTIC velge den siste, dvs. den som er kodet med *høyest* numerisk (eller alfanumerisk) kode, i dette tilfellet kategori 6. En kan imidlertid spesifisere hvilken kategori en vil ha som referansekategori og få den kontroll en ønsker.

4.2. Sammenheng med tabellanalyse

I det følgende skal vi se på et eksempel med samme responsvariabel `deltok` som i eksempelet med 2x2 tabellen, men med en annen forklaringsvariabel, `klasse2`, som er definert i tabellen nedenfor. Før vi går løs på å se på denne tabellen fra et logistisk regresjonssynspunkt skal vi imidlertid studere noe utskrift fra PROC FREQ som også kommer igjen i PROG LOGISTIC. De størrelsene som beregnes lar seg imidlertid best beskrive på grunnlag av PROC FREQ utskriften. Tabellen nedenfor er produsert ved linjene

```
PROC FREQ DATA=lastebilutvalg;
  TABLES deltok*klasse2 / NOROW NOCOL NOPERCENT EXPECTED CHISQ;
RUN;
```

```
deltok(Deltatt i undersøkelsen(0=Nei,1=Ja))
          klasse2(2-7:3,5-8t,8-13t,13-t,trekkb,tankb,andre)
Frequency
Expected  | 2        | 3        | 4        | 5        | 6        | 7        | Total
----------+----------+----------+----------+----------+----------+----------+------
        0 |       30 |       40 |      100 |      112 |       14 |       74 |   370
          |   28.358 |   43.324 |   119.39 |   93.963 |   19.693 |   65.268 |
----------+----------+----------+----------+----------+----------+----------+------
        1 |      222 |      345 |      961 |      723 |      161 |      506 |  2918
          |   223.64 |   341.68 |   941.61 |   741.04 |   155.31 |   514.73 |
----------+----------+----------+----------+----------+----------+----------+------
Total     |      252 |      385 |     1061 |      835 |      175 |      580 |  3288
```

```
Statistics for Table of deltok by klasse2

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     5     11.0169    0.0510
Likelihood Ratio Chi-Square    5     11.1421    0.0486
```

Ved å spesifisere opsjonene `NOROW NOCOL NOPERCENT EXPECTED` har jeg i tillegg til cellefrekvensene fått *forventede* cellefrekvenser på basis av en modell for uavhengighet mellom radvariabel og kolonnevariabel i stedet for rad- og kolonneprosenteringer. Disse er relevante for beregningen av `Chi-Square` og `Likelihood Ratio Chi-Square` statistikkene som altså går igjen i PROC LOGISTIC.

De forventede cellefrekvensene i en uavhengighetsmodell er beregnet ut fra rad- og kolonnemarginaler n_{i+} og n_{+j} som $\hat{m}_{ij} = n_{i+}n_{+j}/n$ der n = tabelltotalen (3288). For celle (1,4) blir dette $\hat{m}_{14} =$ `1061·2918/3288 = 941,61.` `Chi-Square`, eller Pearsons kji-kvadrattest eller X^2 som den også kalles, beregnes som

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} = 11{,}0169 \qquad (4.3)$$

`Likelihood Ratio Chi-Square`, eller G^2 som den gjerne kalles, beregnes som

$$G^2 = \sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{m}_{ij}} = 11{,}1421 \qquad (4.4)$$

De to størrelsene er asymptotisk ekvivalente og kjikvadratfordelte med et antall *frihetsgrader* som er lik forskjellen mellom antall parametre som kreves for å beskrive den fulle tabellen og antall parametre som kreves for å beskrive uavhengighetstabellen. For å beskrive den fulle tabellen kreves like mange parametre som det er celler i tabellen, nemlig $I \cdot J$ hvis tabellen har I rader og J kolonner. For å beskrive uavhengighetsmodellen kreves $I + J - 1$ (= 2 + 6 - 1 = 7) parametre. Differensen blir $IJ - (I + J - 1) = (I - 1)(J - 1) = (2 - 1)(6 - 1) = 5$ i vårt eksempel. Testen viser i dette tilfellet at uavhengighetsmodellen er på grensen til å bli forkastet med et vanlig 5% signifikansnivå.

Kjøring av den logistiske regresjonsmodellen på grunnlag av "individ"datasett `lastebilutvalg` gjøres med følgende kommandoer.

```
PROC LOGISTIC DATA=lastebilutvalg;
  CLASS deltok (DESC) nasj3 region klasse2 (REF='7') ald_kl kj_art2
                                                    / PARAM = REF;
  MODEL deltok = klasse2;
RUN;
```

Opsjonen `(DESC)` er her nødvendig fordi jeg ønsker å modellere sannsynligheten for å svare som er kodet med verdien 1. Det betyr at ikke-svare, som er kodet med 0, blir valgt som referansekategori for responsvariabelen. I utskriften noteres dette med teksten `Probability modeled is deltok=1`. Vi kunne også ha brukt opsjonen `(EVENT=LAST)` (eller `='1'`) eller opsjonen `(REF=FIRST)` (eller `='0'`) eller en opsjon som heter `(ORDER=)`. Uten en av disse opsjonene vil PROC LOGISTIC velge siste kategori, `'1'`, som referanse og altså modellere sannsynligheten for å ikke svare. De to modusene for responsvariabelen er modellmessig ekvivalente, og forskjellen i utskriften vil bare bli at alle parameterestimater skifter fortegn.

`CLASS` kommandoen er nødvendig for å spesifisere alle kategorivariable som *kan* inngå i modellen. Her er det listet opp flere slike variable enn dem som brukes i den aktuelle kjøringen. Opsjonen `PARAM = REF` sier at jeg ønsker å bruke referansekategoriparametrisering. `REF='7'` betyr at kategorien som er kodet `'7'` skal brukes som referansekategori for `klasse2`. (Variabelen er kodet `'2'–'7'.`)

Når alle forklaringsvariablene er kategoriske kan data reduseres til en tabell. PROC FREQ kan lage slike tabeller som så kan leses av LOGISTIC. Datasettet som PROC FREQ vil lage i dette tilfellet får utseendet

Obs	deltok	klasse2	COUNT	PERCENT
1	0	2	30	0.9124
2	0	3	40	1.2165
3	0	4	100	3.0414
4	0	5	112	3.4063
5	0	6	14	0.4258
6	0	7	74	2.2506
7	1	2	222	6.7518
8	1	3	345	10.4927
9	1	4	961	29.2275
10	1	5	723	21.9891
11	1	6	161	4.8966
12	1	7	506	15.3893

For å kjøre modellen fra denne tabellen må kommandoen `WEIGHT = count;` spesifiseres i tillegg. LOGISTIC vil ellers sette `WEIGHT` lik 1 og tro at datasettet er individdata.

Formen på `MODEL` kommandoen brukt ovenfor kalles *Single-trial*. Hvis data er på formen

Obs	klasse2	Ja	I_alt
1	2	222	252
2	3	345	375
3	4	961	1061
4	5	723	835
5	6	161	175
6	7	506	580

kan en bruke formen *Event-trial.* En vil da skrive `MODEL` kommandoen på formen

```
  MODEL Ja/I_alt = klasse2;
```

Event-trial ble brukt i analysen av de grupperte data i tabell 5 og programmet for dette er gjengitt i avsnitt 12.1.3. Bortsett fra dette vil ikke Event-trial bli demonstrert i kurset.
Vi skal nå se på utskriften fra PROC LOGISTIC. Vi ser først på kjøring *uten* `PARAM = REF`.

4.3. Utskrift av kjøringer

4.3.1. Kjøring uten `PARAM = REF`.

Den aller første delen av utskriften bør være grei. De første spørsmålene for den uinvidde kommer med `Class Level Information` tabellen

```
The LOGISTIC Procedure
                                    Model Information

Data Set                      WORK.LASTEBILUTVALG
Response Variable             deltok                 Deltatt i undersøkelsen(0=Nei,1=Ja)
Number of Response Levels     2
Number of Observations        3288
Model                         binary logit
Optimization Technique        Fisher's scoring

          Response Profile

 Ordered                      Total
   Value       deltok     Frequency

       1            1          2918
       2            0           370

Probability modeled is deltok=1.

              Class Level Information

                              Design Variables
Class       Value      1      2      3      4      5

klasse2     2          1      0      0      0      0
            3          0      1      0      0      0
            4          0      0      1      0      0
            5          0      0      0      1      0
            6          0      0      0      0      1
            7         -1     -1     -1     -1     -1
```

Class Level Information matrisen reflekterer valget av parametrisering, som her er sum-til-null eller "effect coding" som SAS kaller det. Den siste linjen er koeffissienter for $\beta_2,\ldots,\beta_6$ som forteller at $\beta_7 = -\beta_2 - \cdots - \beta_6$. Denne linjen uttrykker at vi bruker sum-til-null parametrisering (effect coding) og er et direkte resultat av at `PARAM = REF` ikke brukes.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261 = -2 Log L + 2·1	2314.119 = -2 Log L + 2·6
SC	2321.359 = -2 Log L + 1·Log 3288	2350.707 = -2 Log L + 6·Log 3288
-2 Log L	2313.261	2302.119

-2 Log L er –2 ganger logaritmen til *likelihoodfunksjonen* til modellen, beregnet med de verdier av parametrene (α og β-ene) som er blitt estimert i kjøringen. Det er de verdiene som også maksimerer likelihoodfunksjonen $L(\alpha, \beta)$ med det gitte datasettet. Det er kriteriefunksjonen som brukes for å estimere parametrene. Den uttrykker sannsynligheten for å observere de data som faktisk er observert og benyttet i estimeringen, som funksjon av de ukjente parametrene i modellen. Estimeringsmetoden, Maximum Likelihood (Sannsynlighetsmaksimering på norsk) velger de verdier av α og β som gjør de observerte data mest sannsynlige.

AIC (Akaikes Information Criterium) og SC (Schwartz Criterion) er størrelser som brukes ved valg av modell for å sammenligne flere alternative modeller. Blant flere alternative modeller vil disse kriteriene foreslå den med *lavest* AIC eller SC. De er begge beregnet med utgangspunkt i –2 Log L som demonstrert i utskriften. 6-tallet i beregningene av AIC og SC for "Intercept and Covariates" er antall parametre i modellen, α pluss 6-1=5 selvstendige β-er. I modellen kalt "Intercept only" er bare α med. SC kalles (til forvirring) i store deler av litteraturen for BIC (Bayesian Information Criterion). Enkelt skal det ikke være. Merk ellers at i eksempelet anbefaler SC sterkt modellen med bare konstantleddet mens AIC anbefaler så vidt å ha med `klasse2`.

Begrunnelsen for AIC og SC er å straffe valg av modeller med for mange parametre. De tradisjonelle metodene for valg av modell basert på –2 Log L og signifikanstester som vi skal se på siden, har en tendens til å velge ut modeller som har for mange parametre og som derfor er for komplekse. Dette er særlig tilfelle hvis det er mye data å estimere modellen med. Jo flere parametre, jo bedre tilpassing vil en kunne få mellom modell og data, og jo *lavere* vil –2 Log L bli. Men tillegget som gis i AIC og SC øker med økende antall parametre. SC øker også med antall observasjoner som er til rådighet. AIC og SC vil derfor velge enklere modeller enn metoder som bare er basert på –2 Log L.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.1421 = 2313.261 - 2302.119 = G^2	5 = 6 - 1	0.0486
Score	11.0169 = X^2, Pearsons kji-kvadrat	5	0.0510
Wald	10.9104 Beregnes ikke fra -2 Log L	5	0.0532

`Wald` beregnes på grunnlag av de estimerte regresjonsparametrene og den estimerte matrisen av kovarianser mellom dem. Detaljer vil ikke bli gitt her. De tre testene ovenfor er alle tester av hypotesen "*Alle* β-*ene er 0*" mot at minst en β ikke er det. I den situasjonen at forklaringsvariabelen, eller alle forklaringsvariablene er nominale kategoriske variable, er dette det samme som å teste om det er uavhengighet mellom rader og kolonner i tabellen. Merk at `Likelihood Ratio` blir akkurat det samme som G^2 beregnet i (4.4). `Score` er en statistikk som beregnes på grunnlag av den deriverte av likelihood funksjonen. Den blir i dette enkle tilfellet identisk med X^2 beregnet i (4.3). `Wald` beregnes på grunnlag av en tredje formel basert på $\hat{\beta}$-ene og deres kovariansmatriser. Alle tre testene er asymptotisk kji-kvadratfordelte med (i dette tilfellet) 5 frihetsgrader.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
klasse2	5	10.9104	0.0532

Denne beregnes bare for kategoriske forklaringsvariable og er det samme som `Wald` ovenfor.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.1081	0.0708	887.7520	<.0001
klasse2	2	1	(0.1066 / 0.1739)²	=	0.3761	0.5397
klasse2	3	1	0.0466	0.1536	0.0918	0.7619
klasse2	4	1	0.1547	0.1112	1.9350	0.1642
klasse2	5	1	-0.2432	0.1090	4.9785	0.0257
klasse2	6	1	0.3342	0.2383	1.9679	0.1607

Denne tabellen gir parameterestimatene $\hat{\beta}_j, j = 2,\ldots,6$ og deres standardfeil. `Wald Chi-Square` er `(Estimate/Std Error)²` som er (asymptotisk) kji-kvadratfordelt med 1 frihetsgrad for hver parameter. Merk at det ikke er noen linje for `klasse2 7`. Dette er fordi β_7 kan regnes ut som minus summen av de øvrige β-ene. Det gir β_7 = -0,1857. Jeg synes likevel det er en svakhet ved utskriften at den ikke er med. Ved sum-til-null parametrisering er β_7 ikke i en særklasse i forhold til de andre β-ene.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
klasse2 2 vs 7	1.082	0.688	1.702
klasse2 3 vs 7	1.261	0.839	1.897
klasse2 4 vs 7	1.405	1.021	1.934
klasse2 5 vs 7	0.944	0.689	1.293
klasse2 6 vs 7	1.682	0.925	3.059

De estimerte odds-ratene kan beregnes som $e^{\hat{\beta}_j - \hat{\beta}_7}$ for hver av $\hat{\beta}$-ene i `Analysis of Maximum Likelihood Estimates` tabellen. For eksempel er $e^{\hat{\beta}_3 - \hat{\beta}_7} = e^{0,0466-(-0,1857)} = e^{0,2323} = 1,261$. Tolkningen av dette estimatet er "*odds for å svare i undersøkelsen hvis at klasse2=3 er estimert til* 1,261 *ganger tilsvarende odds hvis klasse2=7*".

4.3.2. Kjøring med `PARAM = REF.`

Det er to forskjeller i utskriften i forhold til kjøring *uten* `PARAM = REF.` Den ene er i er matrisen `Class Level Information` hvor alle tallene i siste linje blir 0 i stedet for –1. Dette reflekterer at klasse2=7 er brukt som referansekategori og at β_7 derfor er satt til 0.

Class Level Information

Class	Value	Design Variables 1	2	3	4	5
klasse2	2	1	0	0	0	0
	3	0	1	0	0	0
	4	0	0	1	0	0
	5	0	0	0	1	0
	6	0	0	0	0	1
	7	0	0	0	0	0

`Analysis of Maximum Likelihood Estimates` tabellen, som nå blir seende slik ut:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.9225	0.1245	238.6020	<.0001
klasse2	2	1	0.0790	0.2309	0.1171	0.7322
klasse2	3	1	0.2322	0.2083	1.2426	0.2650
klasse2	4	1	0.3403	0.1629	4.3658	0.0367
klasse2	5	1	-0.0576	0.1606	0.1284	0.7201
klasse2	6	1	0.5199	0.3052	2.9021	0.0885

`Ved bruk av PARAM = REF` er det ikke så farlig at `klasse2 7` ikke er med siden den likevel vil være 0 . Hvis vi kaller sum-til-null β-ene β_j^E (*E* for "Effect") og referansekategori β-ene for β_j^R vil en kunne se sammenhengen $\beta_j^R = \beta_j^E - \beta_7^E$. Referansekategoriparametrisering står derfor mer direkte i sammenheng med `Odds Ratio Estimates` tabellen i avsnitt 4.3.1. For eksempel er $e^{\beta_3^E} = e^{0,2322} = 1,261$. Dette letter tolkningen av parametrene. Jeg vil derfor anbefale at `PARAM = REF` blir benyttet når kategoriske forklaringsvariable er med.

Tabellen nedenfor sammenfatter og demonstrerer sammenhengen mellom tallene i `Analysis of Maximum Likelihood Estimates` tabellene og logitene i tabellen i begge de to tilfellene. Linjen "Andre" er den som ikke var med i utskriftene fra LOGISTIC.

Tabell 4.1. Odds og logistisk regresjonsparametre i modellen `deltok = klasse2`

	Deltatt		I alt		Logit		PROC LOGISTIC		
	Ja	Nei	n_{i+}, i=2,...,7	ja/nei	log (ja/nei)	=	*Single trial* uten `PARAM = REF`	=	*Event-trial* eller `PARAM = REF`
							$\alpha + \beta_k$	=	$\alpha + \beta_k$
3,5-8 tonn	222	30	252	7,400	2,0015	=	2,1081 – 0,1066	=	1,9225 + 0,0790
8-13 tonn	345	40	385	8,625	2,1547	=	2,1081 + 0,0466	=	1,9225 + 0,2322
13 ^ tonn	961	100	1061	9,610	2,2628	=	2,1081 + 0,1547	=	1,9225 + 0,3403
trekkbiler	723	112	835	6,455	1,8649	=	2,1081 – 0,2432	=	1,9225 – 0,0576
Tankbiler	161	14	175	11,50	2,4423	=	2,1081 + 0,3342	=	1,9225 + 0,5199
Andre	506	74	580	6,838	1,9225	=	2,1081 – 0,1857	=	1,9225 ± 0,0000
n_{+j}, j=1,2	2918	370	3288	7,886	2,0651				

5. Modellenes tilpasning til data

Testene som presenteres i utskriftene i avsnitt 4 indikerer om forklaringsvariabelen/variablene som brukes i modellen har noen betydning for å forklare noe av den variasjon som en kan observere i responsvariabelen. Et annet aspekt er om modellen passer til data i den forstand at den vil reprodusere de originale dataene rimelig bra. Dette er et spørsmål som det er mulig å avgjøre hvis alle forklarings-variablene er kategoriske og hele datasettet kan legges ut som en tabell. Det ideelle vil da være å kunne sammenligne de reproduserte dataene med en såkalt *mettet modell*, men dette er ikke alltid mulig. En mettet modell reproduserer den bakenforliggende tabellen eksakt. Modellen som ble brukt i kapittel 4 er en mettet modell. Hvis en ser på tabell 4.1, vil en se at tallene i kolonnen "log (ja/nei)" lar seg beregne eksakt på grunnlag av de estimerte regresjonsparametrene med eller uten `PARAM = REF`. Sammen med tallene i "I alt" kolonnen vil disse oddsene reprodusere innmaten i tabellen, kolonnene "ja" og "nei" eksakt. Når en kun har en forklaringsvariabel og den er kategorisk vil det alltid være tilfellet. Vi kan derfor se på en modell hvor vi innfører en kategorisk variabel til, `ald_kl` (med 3 kategorier, 1:<= 5 år, 2:6-15 år og 3:16-30 år). Vi kan da sette opp de to modellene

```
I     MODEL deltok = ald_kl klasse2;
```

og

```
II    MODEL deltok = ald_kl klasse2 ald_kl*klasse2;
```

Modell `I` gir

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2271.171
SC	2321.359	2319.955
-2 Log L	2313.261	2255.171

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	58.0902	7	<.0001
Score	58.7238	7	<.0001
Wald	56.7785	7	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ald_kl	2	47.1552	<.0001
klasse2	5	23.5429	0.0003

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.2034	0.1757	46.8968	<.0001
ald_kl	1	1	1.0816	0.1615	44.8277	<.0001
ald_kl	2	1	0.5229	0.1545	11.4491	0.0007
klasse2	2	1	0.3355	0.2376	1.9934	0.1580
klasse2	3	1	0.4620	0.2139	4.6653	0.0308
klasse2	4	1	0.4305	0.1648	6.8187	0.0090
klasse2	5	1	-0.1854	0.1633	1.2889	0.2563
klasse2	6	1	0.5980	0.3076	3.7798	0.0519

Modell `II` gir

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2285.242
SC	2321.359	2395.006
-2 Log L	2313.261	2249.242

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	64.0195	17	<.0001
Score	71.2363	17	<.0001
Wald	65.9954	17	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ald_kl	2	7.1563	0.0279
klasse2	5	6.2104	0.2863
klasse2*ald_kl	10	6.0072	0.8147

Modell II er mettet, fordi den inneholder det høyest mulige interaksjonsleddet `ald_kl*klasse2` men modell I er det ikke. Det er nå mulig (med visse mulige forbehold) å teste modell I mot modell II ved å beregne differensen mellom `-2 Log L` og mellom `DF` for de to modellene:

```
Likelihood ratio Chi-Square = -2 Log L(I) - (-2 Log L(II)) =  2255.171 - 2249.242 = 5.929
                    DF(II - I) = DF(II) - DF(I) = 17 - 7 = 10
                'Pr > ChiSq' = P(ChiSq > 5.929 | DF = 10) = 0.8212
```

Modell I forkastes ikke mot den mettede modellen. Det indikerer at modell I passer godt med data. Denne differensen kalles av noen skribenter for *Deviansen*. For dem som er fortrolig med vanlig lineær regresjon, spiller denne deviansen samme rolle som det som kalles "Residual Sum of Squares". Hvilken modell foretrekker AIC og SC? Det er samme størrelse som G^2 på side 13, men her beregnet i en tre-veis tabell.

Det er verd å merke seg at i modell I, uten interaksjonsleddet , er `ald_kl` har med signifikant **("Pr > ChiSq" <0.0001)** enn `klasse2` **("Pr > ChiSq" = 0.0003)**, men samtidig er `klasse2` sitt forklaringsbidrag blitt mer signifikant etter at `ald_kl` ble bragt inn i modellen enn den var alene **("Pr > ChiSq" = 0.0532)**. I modell II hvor interaksjonsleddet bringes inn mister `klasse2` igjen sin betydning **("Pr > ChiSq" = 0.2863)**.

En fullstendig forståelse av disse sammenligningene ville kreve forståelse av log-lineære modeller og bruk av PROC GENMOD. Hvis det er mange kategoriske forklaringsvariable får en fullstendig tabell fort svært mange celler som vil inneholde tilsvarende få observasjoner hver. Testing mot mettet modell blir da ikke alltid hensiktsmessig. Det kan være mer hensiktsmessig å bare teste mot en mer komplisert modell. Man vil da kunne regne ut og teste på deviansen mellom de to modellene som sammenlignes på samme måte som ovenfor.

Den type testing som er foreslått ovenfor mot mettet eller mer komplisert modell krever at en har faste inndelinger i kategorier. Hvis en av variablene er kontinuerlig vil det ikke uten videre finnes noen slik inndeling. Som omtalt på side 7-8 kan en ikke bare dele inn slik at tilfeldige like verdier som i x-ene i data danner en kategori. Det ville føre til flere grupper jo flere observasjoner. Da ville forutsetninger om frihetsgrader og kji-kvadratfordelinger som testene bygger på ikke holde.

Hosmer og Lemeshow anviser imidlertid en slik automatisk metode som kan brukes approksimativt. Først sorteres observasjonene i stigende rekkefølge etter de estimerte sannsynlighetene for "positiv respons", $\hat{\pi}_i$. På grunnlag av denne sorteringen deles observasjonene inn i ca. 10 intervaller for $\hat{\pi}_i$ med omtrent like mange observasjoner. Med bare én kontinuerlig x-variabel vil dette generere en gruppering av x. Observasjoner med like verdier av $\hat{\pi}_i$ blir ikke splittet på to kategorier selv om det skulle være nødvendig for å få like mange i intervallene. Deretter brukes disse ca.10 kategoriene på samme måte som om de var faste inndelinger og den kategoriske variabelen behandles så som nominal. For parameterestimatene for `ald_kl` i modell I ovenfor, er at $\hat{\beta}_1^A = 1,0815 > \hat{\beta}_2^A = 0,5229 (> \beta_3^A \equiv 0)$, og det ganske signifikant. Toppskrift A erstatter `ald_kl`. Dette indikerer at odds for å svare vil avta med alder (for alle verdier av `klasse2`). Det kan da være fristende å prøve å erstatte `ald_kl` med `alder` som er en kontinuerlig versjon av samme variabelen og kjøre modellen

```
MODEL deltok = alder klasse2 / LACKFIT;
```

Opsjonen `LACKFIT` til slutt ber om at Hosmer-Lemeshows test utføres. Dette gir utskriften

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2269.670
SC	2321.359	2312.356
-2 Log L	2313.261	2255.670

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	57.5912	6	<.0001
Score	59.0282	6	<.0001
Wald	57.2562	6	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
alder	1	47.8274	<.0001
klasse2	5	23.8944	0.0002

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.3673	0.1436	271.7241	<.0001
alder		1	-0.0606	0.00876	47.8274	<.0001
klasse2	2	1	0.4021	0.2399	2.8103	0.0937
klasse2	3	1	0.5085	0.2155	5.5658	0.0183
klasse2	4	1	0.4184	0.1646	6.4612	0.0110
klasse2	5	1	-0.1877	0.1630	1.3269	0.2494
klasse2	6	1	0.5696	0.3072	3.4377	0.0637

Denne modellen har en regresjonsparameter mindre enn da `ald_kl` var med. Spørsmålet er om en kontinuerlig aldersvariabel gir like god tilpasning til data. Her kommer Hosmer-Lemeshow testen inn. `LACKFIT` genererer følgende utskrift:

Partition for the Hosmer and Lemeshow Test

Group	Total	deltok = 1 Observed	deltok = 1 Expected	deltok = 0 Observed	deltok = 0 Expected
1	330	263	261.88	67	68.12
2	287	236	243.24	51	43.76
3	332	286	286.90	46	45.10
4	322	282	282.08	40	39.92
5	280	252	248.05	28	31.95
6	333	310	298.04	23	34.96
7	299	271	270.72	28	28.28
8	303	273	277.78	30	25.22
9	335	308	310.66	27	24.34
10	302	283	282.87	19	19.13
11	165	154	155.74	11	9.26

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
8.2333	9	0.5108

En ser at i dette tilfelle er ca. 10 grupper blitt til 11. `Expected` er beregnet på samme måte som vanlig ved modelltilpassing, nemlig som $n_i\bar{\pi}_i$ der n_i er antall i gruppe i, listet under `Total`, og $\bar{\pi}_i$ er den gjennomsnittlige estimerte sannsynligheten for `deltok = 1` i gruppen. La o_i være det observerte antallet med `deltok = 1` i gruppe i. `Chi-Square` beregnes så etter formelen for Pearsons kji-kvadrattest:

$$X^2_{HL} = \sum_{i=1}^{g(=11)} \frac{(o_i - n_i\bar{\pi}_i)^2}{n_i\bar{\pi}_i(1-\bar{\pi}_i)} \qquad (5.1)$$

9 frihetsgrader er beregnet som 10 = 11 - 1 selvstendige kategorier i inndelingen minus 1 for den estimerte regresjonskoeffisienten $\hat{\beta}^{Alder} (= -0,606)$. Tilpasningen er her god.

I vanlig lineær regresjon er det mer vanlig å måle modellenes tilpassing til data med R^2 som er andelen av variasjonen i Y-ene som forklares av tilpassede verdiene $\hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_k x_k)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \qquad (5.2)$$

Ved å bruke en opsjon som heter `RSQUARE` i `MODEL` kommandoen er det mulig å få ut noen mål som skal illudere R^2. De finnes på side 47 i LOGISTIC manualen og er gjengitt i tekstboksen på neste side.

Det bør her gis noen kommentarer til denne teksten.

1. R^2 som definert i boksen finnes ikke på sidene 208-209 i Cox og Snell. Den finnes ikke i den boken i det hele tatt og jeg vet ikke hvor den er tatt fra.
2. Også i logistisk regresjonsmodulen i SPSS forekommer en størrelse som kalles Nagelkerke R^2. Men det er noe helt annet enn den Nagelkerke R^2 som SAS definerer og ikke meningsfull i en logistisk regresjonssammenheng. "Nagelkerke R^2" refereres i det hele tatt ikke i den litteraturen som er gjengitt i referanselisten bak og må betraktes med skepsis.

R^2 fra vanlig regresjon kan generaliseres til logistisk regresjon. Den mest umiddelbare måten å gjøre det på er

$$R_1^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{\pi}(\mathbf{x}_i))^2}{N\bar{Y}(1-\bar{Y})}. \qquad (5.3)$$

En må her merke seg at når $Y_i = 1$ eller 0 (bare), så blir $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = N\bar{Y}(1-\bar{Y})$ og $\bar{Y}$ andelen med respons 1. Denne størrelsen kan ikke bestilles direkte i PROC LOGISTIC, men ved å skrive ut tilpassede sannsynligheter til datasettet og lage et nytt datasteg kan den beregnes.
En teoretisk mer korrekt måte å beregne en R^2 er å ta utgangspunkt i $-2\log L$. Det kan vises at i vanlig lineær regresjon med normalfordelte restledd, så er

$$R^2 = \frac{-2\log L(\hat{\alpha}) - (-2\log L(\hat{\alpha}, \hat{\boldsymbol{\beta}}))}{-2\log L(\hat{\alpha})} = 1 - \frac{\log L(\hat{\alpha}, \hat{\boldsymbol{\beta}})}{\log L(\hat{\alpha})}. \qquad (5.4)$$

Dette kan generaliseres til logistisk regresjon. Denne størrelsen kan heller ikke bestilles direkte i LOGISTIC, men den er svært enkel å regne ut med kalkulator på grunnlag av det som likevel kommer ut.

Felles for de to siste R^2 målene er at de i logistisk regresjon som regel vil gi en svært lav R^2 i forhold til det som gjerne forekommer i vanlig lineær regresjon. Årsaken til dette er at når responsvariabelen i (binær) logistisk regresjon bare kan anta to verdier, f.eks. 0 eller 1, så vil variasjonen i den (nevneren i R^2) i en viss forstand være maksimal. Hadde man kunne observere Y_i-er mellom 0 og 1 ville de observerte punktene i regresjonsgrafen ligget tettere rundt regresjonskurven slik de kan i vanlig logistisk regresjon. Da ville R^2 blitt høyere. De målene som PROC LOGISTIC kan produsere er forsøk på å bøte på dette.

Generalized Coefficient of Determination

Cox and Snell (1989, pp. 208 -209) propose the following generalization of the coefficient of determination to a more general linear model:

$$R^2 = 1 - \left\{ \frac{L(0)}{L(\hat{\beta})} \right\}^{2/n}$$

where $\boldsymbol{L(0)}$ is the likelihood of the intercept-only model, $L(\hat{\beta})$ is the likelihood of the specified model, and $\boldsymbol{n}$ is the sample size. The quantity $\boldsymbol{R^2}$ achieves a maximum of less than one for discrete models, where the maximum is given by

$$R^2_{\max} = 1 - L(0)^{2/n}$$

Nagelkerke (1991) proposes the following adjusted coefficient, which can achieve a maximum value of one:

$$\tilde{R}^2 = \frac{R^2}{R^2_{\max}}$$

Properties and interpretation of R^2 and $\tilde{R}^2$ are provided in Nagelkerke (1991). In the "Testing Global Null Hypothesis: BETA=0" table, R^2 is labeled as "RSquare" and $\tilde{R}^2$ is labeled as "Max-rescaled Rsquare." Use the RSQUARE option to request R^2 and $\tilde{R}^2$.

6. Metoder for søking etter modell

I den klassiske teorien for statistisk estimering har man (teoretisk) gått ut fra at det eksisterer en "sann" modell og at det er denne en ønsker å estimere. Søking etter "riktig" logistisk regresjonsmodell vil da bestå i å finne de forklaringsvariablene som har innvirkning på responsvariablene og deretter estimere dem best mulig. Det at modellen virkelig er lineær på akkurat logistisk skala er en forutsetning som tas for gitt.

I virkelighetens verden eksisterer ingen "sann" modell, bare modeller som mer eller mindre adekvat beskriver de fenomener en ønsker å studere. Det er ikke sikkert at vi har tilgjengelig akkurat de forklaringsvariablene som egentlig virker, og vi må bruke dem vi har tilgjengelig som mer eller mindre proxy variable. Det er ikke sikkert at de variablene vi har tilgjengelige virker lineært på logistisk skala. For i størst mulig grad å unngå å måtte gjøre antagelser som kanskje ikke holder er det utviklet mer generelle modeller, som f.eks. *generaliserte additive modeller* som kan være av typen

$$\operatorname{logit} \pi(\mathbf{x}) = \alpha + g_1(x_1) + g_2(x_2) + \cdots + g_k(x_k)\,. \qquad (6.1)$$

hvor det ikke gjøres andre antagelser om $g_1, \ldots, g_k$ enn at de er tilstrekkelig glatte, og så lar man data bestemme funksjonsformen. I det videre skal vi imidlertid holde oss innenfor rammene av lineære logistiske modeller.

Hvilken logistisk regresjonsmodell som er best avhenger ikke bare av hvilken som er den beste tilnærmingen til virkeligheten, men også hva en vil med den modellen en ønsker å estimere. Et mulig formål med å søke etter en modell kan være ønsket om best mulig å forstå et fenomen og hvilke faktorer det er som påvirker variasjonen i responsvariabelen. Et annet formål kan være å predikere verdien til responsvariabelen for statistiske enheter hvor bare forklaringsvariabelen er observert. For eksempel kunne man tenke seg å estimere en modell hvor forklaringsvariablene finnes for alle i et

register mens responsvariabelen bare er tilgjengelig i et utvalg. Så ønsker man å bruke en estimert logistisk regresjonsmodell til å anslå sannsynlighetene for 0 eller 1 i responsvariabelen for dem som ikke var med i utvalget. For eksempel kan man ønske å si noe om sannsynligheten for at en person er sysselsatt, arbeidssøker eller utenfor arbeidsstyrken på grunnlag av registerdata. Et steg videre ville være å bruke de estimerte sannsynlighetene til å klassifisere med. Da er det ikke likegyldig om man ønsker at klassifikatoren skal klassifisere flest mulig individer riktig (for eksempel med hensyn på sysselsetting) eller om man ønsker at de klassifiserte verdiene skal følge statistisk fordeling som mest mulig riktig reflekterer den virkelige fordelingen i befolkningen. Dette var et sentralt problem ved utvikingen av de metodene som ble brukt i persondelen av folketellingen 2001, selv om metodene som ble benyttet til dels var andre. Et annet eksempel er bruk av slike modeller for å bestemme diagnosekriterier for sykdommer.

Regresjonsmodeller som skal brukes til prediksjon vil ofte med fordel gjøres enklere med ferre parametre og variable enn dem som bare skal benyttes til "forståelse". Usikkerheten i de estimerte modellparametrene bidrar til usikkerheten i prediksjonene som lages på grunnlag av dem. Og jo flere parametre, jo mer usikkert vil hver av dem bli estimert. De må dele på data. Kvitter man seg med en parameter (og variabel) som spiller liten rolle, kan det bidra til en liten skjevhet i prediksjonene, men kan gi en større gevinst i form av redusert prediksjonsusikkerhet. Her kommer AIC og SC kriteriene inn i forsøk på å skape balanse. Et annet aspekt er at hvis datamengden som er til rådighet for å estimere modellen er stor nok vil selv effekter av variable som har en liten eller ubetydelig innvirkning på responsen bli funnet statistisk signifikante. Statistisk signifikans er ikke det samme som faglig relevans.

I dette kapitlet skal vi ikke ta opp alle de aspektene som er nevnt ovenfor, men begrense oss til det som er tilgjengelig innen rammen av PROC LOGISTIC.

6.1. Ordnede kategorier i forklaringsvariable

Ved bruk av kategoriske forklaringsvariable med ordnede kategorier kan man prøve å behandle disse som en ordinal variabel i stedet for nominal. I eksemplene foran er `ald_kl` en slik variabel. Analyse av parameterestimatene i modell I, `deltok = ald_kl klasse2`, indikerer at en modell hvor denne blir behandlet som ordinal med verdiene 1, 2 og 3 ville fungere godt. Den ville bare ha en regresjonsparameter knyttet til variabelen i stedet for en til hver av de to første kategoriene. For å gjennomføre dette teknisk i PROC LOGISTIC vil det være tilstrekkelig å fjerne variabelen fra listen over CLASS variable. I dette tilfellet hadde vi også den "nesten" kontinuerlige aldersvariabelen `alder` tilgjengelig, og denne fungerte i vårt tilfelle like bra. Slike kontinuerlige versjoner av ordinale diskrete variable er imidlertid ikke alltid tilgjengelig.

6.2. Kollapsing av kategorier i forklaringsvariable

En måte å redusere antall parametre i de logistiske regresjonsmodellene er å slå sammen kategorier i forklaringsvariablene. Hvis to eller flere kategorier synes å ha svært lik effekt på responsen og det ikke er noen apriori grunn til at de skulle ha signifikant ulike effekter, er det grunn til å vurdere sammenslåing. Men før en gjør det bør en teste eksplisitt om effektene av de kategoriene en ønsker å slå sammen kan være like. Hvis slike tester gir svært lav p-verdi bør en revurdere det. I PROC LOGISTIC kan CONTRAST kommandoen brukes til å gjøre slike tester.

En *kontrast* i parametrene er en lineærkombinasjon (veid sum) av dem hvor summen av koeffissientene er 0. Hvis det er f.eks. 6 β-parametre, $\beta_1,\ldots,\beta_6$, er

$$L(\beta_1,\ldots,\beta_6) = c_1\beta_1 + c_2\beta_2 + \cdots + c_6\beta_6 \qquad (6.2)$$

en kontrast hvis $c_1 + c_2 + \cdots + c_6 = 0$. Oppgaven blir å teste om $L(\beta_1, \ldots, \beta_6) = 0$.

I eksempelet vårt er det grunn til å stille spørsmål om β-ene til de tre første kategoriene i `klasse2` egentlig er forskjellige. De tre kategoriene skiller ulike vektklasser av "vanlige" lastebiler, og det er få apriori grunner til å tro at den bakenforliggende svartilbøyeligheten skulle være særlig forskjellig i de tre kategoriene. I den siste modellen som ble estimert i kapittel 5 `(deltok = alder klasse2 / LACKFIT)` ble de tre β-estimatene $\hat{\beta}_2 = 0,4021, \hat{\beta}_3 = 0,5085$ og $\hat{\beta}_4 = 0,4184$.

For å sammenligne disse tre parametrene, dvs. teste hypotesen $\beta_2 = \beta_3 = \beta_4$ trenger vi to kontraster. Disse kan f.eks. formuleres som

$$\begin{aligned} L_1(\beta_2,\ldots,\beta_7) &= \beta_2 - \beta_4 \\ L_2(\beta_2,\ldots,\beta_7) &= \beta_3 - \beta_4 \end{aligned} \qquad (6.3)$$

For å teste om disse kontrastene er 0 kan vi bruke følgende oppsett:

```
PROC LOGISTIC DATA=lastebilutvalg;
  CLASS deltok (DESC) nasj3 region klasse2 (REF='7') kl3 ald_kl kj_art2 /
                                                          PARAM = REF;
  MODEL deltok = alder klasse2;
  CONTRAST 'klasse2=2 mot klasse2=4' klasse2 1 0 -1 0 0;
  CONTRAST 'klasse2=3 mot klasse2=4' klasse2 0 1 -1 0 0;
RUN;
```

`CONTRAST` kommandoen angir først en merkelapp hvor vi kan spesifisere hva kontrasten gjelder, dernest variabelen den angår og til sist $c_2, \ldots, c_6$. Merk at den c_7 ikke angis. Den regnes på grunnlag av de øvrige. Dette er spesielt viktig å være oppmerksom på hvis den siste kategorien skulle inngå i kontrasten med $c_7 \neq 0$. Hvordan `CONTRAST` koeffisientene spesifiseres avhenger også av parametrisering. `CONTRAST` kommandoene genererer så følgende utskift:

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
klasse2=2 mot klasse2=4	1	0.0051	0.9428
klasse2=3 mot klasse2=4	1	0.1985	0.6559

Disse indikerer at ingen av kontrastene er signifikant forskjellig fra 0. Det vil være rimelig å erstatte `klasse2` med en variabel hvor disse kategoriene er slått sammen. En slik variabel er `kl3`. Denne har da kategoriene `(4:Lastebil,5:trekkbil,6:tankbil,7:andre)`. Kjøring av modellen med `kl3` i stedet for `klasse2`,

```
MODEL deltok = alder kl3;
```

gir

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2265.912
SC	2321.359	2296.402
-2 Log L	2313.261	2255.912

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	57.3496	4	<.0001
Score	58.7928	4	<.0001
Wald	57.0300	4	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
alder	1	49.4721	<.0001
kl3	3	23.6855	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.3641	0.1428	274.1693	<.0001
alder		1	-0.0602	0.00856	49.4721	<.0001
kl3	4	1	0.4360	0.1522	8.2095	0.0042
kl3	5	1	-0.1868	0.1629	1.3151	0.2515
kl3	6	1	0.5691	0.3072	3.4332	0.0639

Sammenligninger med tidligere kjøringer viser også at AIC og SC er betydelig redusert som følge av en reduksjon av antallet modellparametre med 2, mens -2 log L er neste uforandret:

```
-2 log L(M2) - (-2 log L(M1)) = 2255.912 - 2255.670 = 0.242,   DF = 2
```

Parameterestimatene er heller ikke nevneverdig påvirket. Man kunne stille spørsmål om flere kategorier kunne slås sammen. Tilsynelatende er verken kategori 5 eller 6 ”signifikante”. Man må da huske på at kji-kvadrattesten i høyre kolonne ikke tester effekten av kategoriene som sådan, men om effektene av dem er forskjellige fra effekten av referansekategorien, kategori 7. Effekten av kategoriene 5 og 6 har motsatt fortegn, og selv om ingen av dem skulle være signifikant forskjellige fra referansekategorien, kan de være signifikant forskjellige fra hverandre. Test av en kontrast gir

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
kl3=5 mot kl3=6	1	6.3599	0.0117

I tillegg kommer faglige vurderinger. Det er ikke urimelig å tro at svartilbøyeligheten blant tankbiler `(kl3 = 6)` vil være høyere enn blant trekkbiler slik som parametrene kan indikere fordi tankbiler er enklere å svare for. De kan bare frakte en type last, og har ofte faste ruter å kjøre.

6.3. Trinnvise prosedyrer

Vi har hittil bare sett på noen få variable som kan inngå. I tillegg til `alder` og `kl3` har vi tidligere sett på type kjøring, kalt `kj_art2.region` med 4 kategorier, er en variabel som også er tilgjengelig, `nasj3` som angir om bilen tilhører firma med såkalt utenlandslisene og den kontinuerlige variabelen `nyttelas`. En gruppert versjon av `nyttelas` ble brukt som en del av `klasse2` for å skille mellom ulike størrelser av vanlige lastebiler. Vi kunne ikke påvise noen egen effekt av

nyttelastkapasiteten på svartilbøyeligheten for disse bilene, og det er heller ingen grunn til at den skal ha slik effekt for andre typer biler.

Til hjelp for å velge mellom modeller har de fleste programmer for logistisk regresjon innlagt muligheten for såkalt trinnvise seleksjonsprosedyrer for valg av variable. Ved bruk av disse prosedyrene legger vi inn i `MODEL` alle de variable som vi betrakter som kandidater til å være forklaringsvariable i den endelige modellen. PROC LOGISTIC sammenligner så modeller ved å ta inn eller hive ut variable.

I PROC LOGISTIC ligger det inne fem ulike metoder for trinnvis regresjon:

NONE: PROC LOGISTIC kjører bare den modellen som er oppgitt i `MODEL`.

FORWARD: Først kjøres modellen med bare konstantledd. Denne testes mot modellen som inneholder alle variablene. Så kjøres `MODEL` med konstantledd og en for en av kandidatvariablene og `Score` chi-square, tilhørende frihetsgrader og *p*-verdi beregnes for hver av dem. Den laveste av disse *p*-verdiene sammenlignes med et testnivå som kan spesifiseres med opsjonen `SLENTRY = .` `SLENTRY = 0.05` er standard hvis ikke noe annet er spesifisert. Hvis denne minste *p*-verdien er mindre enn `SLENTRY`, tas den tilhørende variabelen med som første variabel i modellen. Deretter velges den av de øvrige som sammen med den først valgte variabelen gir den mest signifikante økningen i `Score`, forutsatt at *p*-verdien er mindre enn `SLENTRY`. Slik fortsetter søkingen inntil ingen flere variable kan gi en signifikant økning i `Score`. Metoden går bare en vei. Straks en variabel er inkludert i modellen blir den værende der. Hosmer og Lemeshow (2000, s.118) kommenterer: "More recently Lee and Koval (1997) examined the issue of significance level in forward stepwise regression. The results of this research have shown that the choice of $p_E = 0.05$ (`SLENTRY` , min kommentar) is too stringent, often excluding important variables from the modell. Choosing p_E in the range 0.15 to 0.20 is highly recommended". Dette avhenger imidlertid av hva vi ønsker å bruke den endelige modellen til.

BACKWARD: Denne metoden starter med alle kandidatvariablene og hiver ut den som er *minst* signifikante målt med `Wald` statistikken forutsatt at denne er mindre signifikant enn det som er oppgitt i opsjonen `SLSTAY`. Hvis `SLSTAY` ikke er oppgitt antas den å være 0,05. Metoden fortsetter så med de resterende variablene inntil den ikke lenger kan kaste noen ut. Hosmer og Lemeshow anbefaler også her å bruke `SLSTAY` verdier av størrelsesorden 0,20.

STEPWISE: Denne begynner som FORWARD, men kan kaste ut igjen variable ved å bruke de kriteriene som BACKWARD benytter. Den fortsetter slik inntil den verken kan inkludere en ny eller hive ut en gammel variabel. Metoden kan ende med å oscillere mellom to modeller uten å kunne bestemme seg. PROC LOGISTIC vil oppdage dette og avbryte. Ellers er det mulig å unngå det ved å velge `SLSTAY > SLENTRY`. H&L skriver: "If we do not wish to exclude many variables once they are entered then we might use $p_R = 0.9$ (`SLSTAY`). A more stringent value would be used if a continued 'significant' contribution is required. For example, if we used $p_E = 0.15$ then we might choose $p_R = 0.20$."

Mange program bruker `Likelihood Ratio` både forlengs og baklengs i stedet for `Score` og `Wald` som SAS gjør. `Likelihood Ratio` er fra et teoretisk synspunkt bedre, men er også vesentlig mer beregningsmessig krevende fordi alle modeller som skal sammenlignes da må estimeres. Det samme gjelder hvis vi baserer valget på AIC eller BC, noe som også ville være naturlig. Dette er ikke nødvendig med `Score` og `Wald` og dette er antakelig årsaken til at SAS gjør slik den gjør.

SCORE: Bruker en egen søkemetode for å finne den modellen som har høyest Score Chi Square blant alle modeller med henholdsvis 1, 2, 3 osv. variable. Se for øvrig PROC LOGISTIC manualen.

Vi skal nå se på utskrift fra kjøring av

```
PROC LOGISTIC DATA=lastebilutvalg;
  CLASS deltok (DESC) nasj3 region klasse2 (REF='7') kl3 ald_kl kj_art2 /
                                                          PARAM = REF;
  MODEL deltok = alder kl3 nasj3 region kj_art2
    / SELECTION=STEPWISE SLENTRY=0.15 SLSTAY=0.20 HIERARCHY=SINGLE; RUN;
```

Det vil fremgå av utskriften at `STEPWISE` i dette tilfellet ikke kaster ut noen variable underveis. Utskriften blir derfor identisk lik den vi ville fått fra `FORWARD`.

```
Step  0. Intercept entered:
```

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
82.6929	9	<.0001

En modell med alle variablene ville hatt 10 parametre mot 1 for bare konstantleddet. 10-1=9. Tabellen viser bare at noen av variablene må ha en signifikant effekt i forhold til konstantleddet.

```
Step  1. Effect alder entered:
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2283.669
SC	2321.359	2295.865
-2 Log L	2313.261	2279.669

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.5920	1	<.0001
Score	36.1014	1	<.0001
Wald	35.2565	1	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
49.1872	8	<.0001

Bilens alder er tydeligvis den mest signifikante faktoren når det gjelder å forklare svartilbøyelighet. `Residual Chi-Square Test` viser imidlertid at mye variasjon gjenstår å forklare i forhold til det alle de tilgjengelige variablene kan klare.

```
Step  2. Effect kl3 entered:
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2265.912
SC	2321.359	2296.402
-2 Log L	2313.261	2255.912

Testing Global Null Hypothesis: BETA=0

Sammenligner modell med alder og kl3 mot den med bare konstantledd.

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	57.3496	4	<.0001
Score	58.7928	4	<.0001
Wald	57.0300	4	<.0001

Residual Chi-Square Test

Sammenligner modell med alder og kl3 mot den med alle variablene.

Chi-Square	DF	Pr > ChiSq
25.1647	5	0.0001

`kl3` tas med som andre. Merk imidlertid at SC har steget noe. SC kriteriet ville derfor ikke ha tatt med `kl3` om det hadde kunnet bestemme.

```
Step  3. Effect region entered:
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2256.297
SC	2321.359	2305.082
-2 Log L	2313.261	2240.297

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	72.9638	7	<.0001
Score	74.8377	7	<.0001
Wald	72.0237	7	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
8.8368	2	0.0121

SC fortsetter å stige etter at `region` er tatt med som tredje variabel, mens AIC fortsetter å synke.

```
Step  4. Effect nasj3 entered:
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2249.360
SC	2321.359	2304.242
-2 Log L	2313.261	2231.360

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	81.9011	8	<.0001
Score	82.6763	8	<.0001
Wald	79.3504	8	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.0185	1	0.8919

Det er nå svært god tilpassing i forhold til modellen med alle variablene. SC har sunket litt igjen.

```
NOTE: No effects for the model in Step 4 are removed.
NOTE: No (additional) effects met the 0.15 significance level for entry into the model.
```

Variabelen `kj_art2` tas ikke med. Det ser ikke ut til at den har noen betydning. PROC LOGISTIC oppsummerer nå med

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Forlengs Score Chi-Square	Baklengs Wald Chi-Square	Pr > ChiSq	Variable Label
1	alder		1	1	36.1014	.	<.0001	Kjøretøyets alder
2	kl3		3	2	24.1339	.	<.0001	4-7:lastebiler, trekkb,tankb,andre
3	region		3	3	16.3373	.	0.0010	1=fylke 1-3, 2=f4-8, 3=f9-15, 4=f16-20
4	nasj3		1	4	8.8209	.	0.0030	1=NS, 2=IS, (som nasj2, men 4 omkodet til 1

For den modellen som til slutt blir valgt produseres så den vanlige utskriften:

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
alder	1	41.2588	<.0001
kl3	3	23.9294	<.0001
nasj3	1	8.7610	0.0031
region	3	15.0493	0.0018

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.7953	0.1945	206.4672	<.0001
alder		1	-0.0574	0.00894	41.2588	<.0001
kl3	4	1	0.4038	0.1532	6.9501	0.0084
kl3	5	1	-0.2469	0.1665	2.2001	0.1380
kl3	6	1	0.5519	0.3083	3.2037	0.0735
nasj3	1	1	-0.3756	0.1269	8.7610	0.0031
region	1	1	-0.5239	0.1576	11.0462	0.0009
region	2	1	-0.0340	0.1672	0.0415	0.8386
region	3	1	-0.1310	0.1661	0.6217	0.4304

Variabelen `region` har fire kategorier. Det kan synes som om svartilbøyeligheten i region 2 og 3 ikke er signifikant forskjellig fra den i region 4, mens region 1 skiller seg ut. *p*-verdiene som kommer ut for disse under `Pr > ChiSq` er for tester av kontrastene $\beta_2^R - \beta_4^R = 0$ og $\beta_3^R - \beta_4^R = 0$ og derved det samme som vi ville få ved å bruke

```
CONTRAST 'Region 2 mot 4' region 0 1 0;
CONTRAST 'Region 3 mot 4' region 0 0 1;
```

(Merk igjen at c_4 ikke spesifiseres. Den er –1 i begge kontrastene og det regnes ut automatisk.) Kjøring av den valgte modellen med disse kontrastene gir utskriften

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Region 2 mot 4	1	0.0415	0.8385
Region 3 mot 4	1	0.6215	0.4305

Hvis vi så omkoder region til en ny variabel regi2 hvor regionene 2,3 og 4 er slått sammen og kjører på nytt får vi

Model Fit Statistics

Criterion	Intercept Only	med regi2 Intercept and Covariates	med region Intercept and Covariates
AIC	2315.261	2246.029	< 2249.360
SC	2321.359	2288.715	< 2304.242
-2 Log L	2313.261	2232.029	> 2231.360

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	81.2322	6	<.0001
Score	82.1074	6	<.0001
Wald	78.8293	6	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
alder	1	40.6597	<.0001
kl3	3	24.0894	<.0001
nasj3	1	8.9099	0.0028
regi2	1	14.4416	0.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.7403	0.1679	266.3947	<.0001
alder		1	-0.0568	0.00890	40.6597	<.0001
kl3	4	1	0.3987	0.1530	6.7917	0.0092
kl3	5	1	-0.2546	0.1661	2.3481	0.1254
kl3	6	1	0.5538	0.3081	3.2311	0.0723
nasj3	1	1	-0.3787	0.1269	8.9099	0.0028
regi2	1	1	-0.4668	0.1228	14.4416	0.0001

Denne siste modellen slår alle de andre både når vurdert fra AIC og SC. Økningen i –2Log L er på 0,669 som ikke er en signifikant økning med 2 frihetsgrader (= 9 parametre – 7 parametre). Den siste modellen kan bli stående.

7. Interaksjonseffekter

Med unntak av den mettede modell II med to forklaringsvariable i kapittel 5 har vi ikke studert logistiske regresjonsmodeller med 2. ordens effekter. Vi skal derfor se på en modell av typen

$$\operatorname{logit}\pi(x_1,x_2)=\log\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}=\alpha+\beta_1x_1+\beta_2x_2+\beta_{12}x_1x_2\,. \qquad (7.1)$$

Hvis vi nå øker f.eks. x_2 med en enhet, får vi

$$\operatorname{logit}\pi(x_1,x_2+1)=\log\frac{\pi(x_1,x_2+1)}{1-\pi(x_1,x_2+1)}=\alpha+\beta_1x_1+\beta_2(x_2+1)+\beta_{12}x_1(x_2+1) \qquad (7.2)$$

Tar vi differensen (7.2) - (7.1) får vi

$$\operatorname{logit}\pi(x_1,x_2+1)-\operatorname{logit}\pi(x_1,x_2)=\log\frac{\pi(x_1,x_2+1)/(1-\pi(x_1,x_2+1))}{\pi(x_1,x_2)/(1-\pi(x_1,x_2))}=\beta_2+\beta_{12}x_1 \quad (7.3)$$

Med andre ord: Effekten av x_2, endringen i odds og log odds som følge av en økning i x_2 med 1, er nå avhengig av verdien på x_1. Tilsvarende blir

$$\operatorname{logit}\pi(x_1+1,x_2)-\operatorname{logit}\pi(x_1,x_2)=\log\frac{\pi(x_1+1,x_2)/(1-\pi(x_1+1,x_2))}{\pi(x_1,x_2)/(1-\pi(x_1,x_2))}=\beta_1+\beta_{12}x_2\,, \quad (7.4)$$

effekten av av x_1, nå avhengig av verdien på x_2.

Hvis vi ser på forskjellen i effekten av x_2 for to verdier av x_1 som adskiller seg med 1, får vi

$$
\begin{aligned}
&(\operatorname{logit}\pi(x_1+1,x_2+1)-\operatorname{logit}\pi(x_1+1,x_2))-(\operatorname{logit}\pi(x_1,x_2+1)-\operatorname{logit}\pi(x_1,x_2))\\
&=\log\left(\frac{\pi(x_1+1,x_2+1)/(1-\pi(x_1+1,x_2+1))}{\pi(x_1+1,x_2)/(1-\pi(x_1+1,x_2))}\Big/\frac{\pi(x_1,x_2+1)/(1-\pi(x_1,x_2+1))}{\pi(x_1,x_2)/(1-\pi(x_1,x_2))}\right)\\
&=(\beta_2+\beta_{12}(x_1+1))-(\beta_2+\beta_{12}x_1)=\beta_{12}
\end{aligned}
$$

som gir en tolkning av β_{12}. Ser en på den midterste linjen ser en at β_{12} er logaritmen til forholdet mellom to oddsforhold.

Noe mer oversiktlig blir dette hvis vi tenker oss at x_1 og x_2 er rene 0-1 variable, en situasjon som kan beskrives ved en 2 x 2 x 2 tabell. Effekten av å endre x_2 fra 0 til 1 når $x_1 = 0$ blir

$$
\operatorname{logit}\pi(0,1)-\operatorname{logit}\pi(0,0)=\log\frac{\pi(0,1)/(1-\pi(0,1))}{\pi(0,0)/(1-\pi(0,0))}=\beta_2+\beta_{12}\cdot 0=\beta_2\,,
$$

mens den tilsvarende effekten når $x_1 = 1$ blir

$$
\operatorname{logit}\pi(1,1)-\operatorname{logit}\pi(1,0)=\log\frac{\pi(1,1)/(1-\pi(1,1))}{\pi(1,0)/(1-\pi(1,0))}=\beta_2+\beta_{12}\cdot 1=\beta_2+\beta_{12}\,.
$$

Med andre ord: β_{12} er forskjellen i effekten av x_2 mellom de to nivåene av x_1. Helt tilsvarende regninger viser at dette er det samme som forskjellen i effekten av x_1 mellom de to nivåene av x_2.

Som eksempel kan vi nå prøve å legge et slikt interaksjonsledd til modellen vi kom frem til i kapittel 6. Vi kan for eksempel se om det er grunnlag for å si at effekten av alder er forskjellig for de to mulige nivåene 1 og 2 av `nasj3`:

```
MODEL deltok = alder kl3 nasj3 regi2 alder*nasj3;
```

Utskrift:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2315.261	2243.653
SC	2321.359	2292.438
-2 Log L	2313.261	2227.653

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	85.6078	7	<.0001
Score	84.1320	7	<.0001
Wald	80.6088	7	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
alder	1	23.6720	<.0001
kl3	3	25.1439	<.0001
nasj3	1	12.2452	0.0005
regi2	1	13.5836	0.0002
alder*nasj3	1	4.5010	0.0339

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.9776	0.2058	209.4146	<.0001
alder		1	-0.0931	0.0191	23.6720	<.0001
kl3	4	1	0.3858	0.1530	6.3617	0.0117
kl3	5	1	-0.2851	0.1668	2.9201	0.0875
kl3	6	1	0.5551	0.3083	3.2421	0.0718
nasj3	1	1	-0.6927	0.1979	12.2452	0.0005
regi2	1	1	-0.4538	0.1231	13.5836	0.0002
alder*nasj3	1	1	0.0447	0.0211	4.5010	0.0339

Estimatet for `alder` endrer seg nå fra –0,0568 til –0,0931. Dette er da den estimerte effekten av `alder` når `nasj3 = '2'`, eierfirma med utenlandslisens, som er referansekategorien. Den estimerte effekten av `alder` når `nasj3 = '1'` blir –0,0931 + 0,0447 = -0,0484. Utskriften viser en mulig signifikant effekt av interaksjonsleddet: tilbøyeligheten til å svare i undersøkelsen avtar raskere med alder for biler som tilhører firma med utenlandslisens enn for biler som ikke gjør det. Estimatet for `alder` uten interaksjonsledd, –0,0568, ligger betydelig nærmere -0,0484 enn –0,0931. Dette kan skyldes at det er en betydelig høyere andel eldre biler blant dem med `nasj3 = '1'` enn blant dem med `nasj3 = '2'`. Disse kan ha hatt stor innflytelse på estimatet.

I logistisk regresjon er det vanlig å kreve at dersom en interaksjon som `alder*nasj3` skal være med i modellen, så skal `alder` og `nasj3` også være det. Tilsvarende, hvis tredje ordens ledd som `alder*nasj3*kl3` hadde vært med skulle alle tre 2. ordens ledd som består av to av de tre variablene være med og dessuten alle tre variablene hver for seg. Et slikt system av modeller kalles *hierarkisk*. En modell der f.eks. `alder` ikke er med, men `nasj3` og `alder*nasj3` er med vil måtte tolkes slik at `alder` ikke har noen effekt når `nasj3` er på sitt referansekategorinivå ('2') men at alder har effekt når `nasj3` er '1'. Hvis både `alder` og `nasj3` mangler, må tolkningen bli at `nasj3` i tillegg ikke har noen egen effekt når (den numeriske) variabelen `alder=0` . Slike situasjoner kan selvfølgelig tenkes å oppstå, men vil ofte oppfattes som kunstige. I trinnvise regresjonsprosedyrer i SAS tas det ikke automatisk hensyn til at modellen skal være hierarkisk oppbygd. I den trinnvise prosedyren kan det derfor hende at et interaksjonsledd i listen i `MODEL` kommandoen blir tatt inn før en eller noen av de variablene som interaksjonsleddet består av. Hvis dette ikke er ønskelig , er det mulig å tvinge den trinnvise prosedyren til å ta inn variablene i hierarkisk rekkefølge. Dette kan gjøres ved opsjonen `HIERARCHY = keyword` (eller forkortet til `HIER=keyword`) i `MODEL` kommandoen. Som `keyword` kan man velge

`NONE`, som betyr at `HIERARCHY` er satt ut av effekt,
`SINGLE`, som sier at bare en effekt kan tas inn eller hives ut om gangen. Dette er standard.
`SINGLECLASS`, som er det samme som `SINGLE`, men bare anvendt på `CLASS` variable,
`MULTIPLE`, som tillater å ta inn eller hive ut flere effekter av gangen, og
`MULTIPLECLASS` som er det samme som `MULTIPLE`, men bare anvendt på `CLASS` variable.

Jeg vil anbefale bruk av SINGLE, ev. SINGLECLASS. Det gir bedre kontroll over hva som foregår. HIERARCHY=keyword virker bare hvis også SELECTION = brukes.

En annen opsjon som kan brukes sammen med SELECTION er STOPRES (SR). Denne krever at Chi Square som brukes av FORWARD også brukes av BACKWARD og STEPWISE i stedet for Wald. Tas alle opsjoner (unntatt STOPRES) med (med sine standard verdier som ble brukt) vil MODEL kommandoen for den trinnvise regresjonen i kapittel 6 se slik ut:

```
MODEL deltok = alder kl3 nasj3 region kj_art2 nasj3*region nasj3*kl3
region*kl3 nasj3*ald_kl
    / SELECTION=STEPWISE SLENTRY=0.15 SLSTAY=0.20 HIERARCHY=SINGLE;
```

I eksempelet i kapittel 6 ble alle variable tatt med enkeltvis og seleksjonen stoppet før noen interaksjonsledd kom med. Bruk av HIERARCHY var derfor ikke nødvendig.

8. Proporsjonal odds modell

Når responsvariabelen har flere enn to kategorier og disse er ordinale slik som noen av variablene listet i tabell 2 på side 3, vil vi ofte ønske å kunne ta hensyn til ordinaliteten i analysen av variablene. Det vil kunne gi mer relevante analyser med klarere resultat og ferre parametre.

Det finnes mange måter å modellere ordinalitet i responsvariabelen. Innen rammen av logistisk regresjon skiller modellene seg med hensyn på hvordan man lager odds og log odds. PROC LOGISTIC har bare én slik modell innebygd, proporsjonal odds modellen. Dersom responsvariabelen har mer enn to kategorier i data vil LOGISTIC som standard anta at kategoriene er ordinale og bruke proporsjonal odds modellen. Det er imidlertid mulig å kreve at responskategoriene skal betraktes som nominale ved å spesifisere LINK = glogit som opsjon i MODEL kommandoen. Se Logistic Procedure Update manualen.

Anta at variabelen Y kan anta en av verdiene $j = 1, ..., J$ med sannsynligheter $\pi_1, ..., \pi_J$. Vi lager oddsene

$$\text{logit}\, P(Y \le j) = \log \frac{P(Y \le j)}{1 - P(Y \le j)} = \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} \right), \quad j = 1, \ldots, J \qquad (8.1)$$

Dette kalles *kumulative* logits. Vi kan tenke oss en logit modell for $P(Y \le j \mid \boldsymbol{x})$ beskrevet som

$$\text{logit}\, P(Y \le j \mid \boldsymbol{x}) = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_K x_K \qquad (8.2)$$

Merk at j bare hekter på β_0, ikke på de andre β-ene. Det er fordi vi må ha $P(Y \le 1 \mid \boldsymbol{x}) \le P(Y \le 2 \mid \boldsymbol{x}) \le \ldots \le P(Y \le J \mid \boldsymbol{x}) = 1$ for alle mulige verdier av $\boldsymbol{x}$ og følgelig også

$$\beta_{0j} + \beta_{1j} x_1 + \cdots + \beta_{Kj} x_K \le \beta_{0l} + \beta_{1l} x_1 + \cdots + \beta_{Kl} x_K \text{ når } j < l\,. \qquad (8.3)$$

for *alle mulige* verdier av $\boldsymbol{x}$. Dette er bare mulig hvis regresjonslinjene er parallelle, som betyr at alle β-ene unntatt β_0 må være like. Derfor kalles dette også ”parallel lines regression”.

Som et eksempel kan vi se på bønders yrkestilknytning til gården. Vi kan dele bøndene inn i tre kategorier:

1. De som bare har inntekt fra gården
2. De som har gården som hovedyrke men har bijobb ved siden av, og
3. De som har hovedyrket utenfor gården og driver gården ved siden av.

Torkil Løwe har analysert den modellen vi skal se på nedenfor, og resultatene er publisert i Økonomiske Analyser 6/2003. Som kovariater har vi gårdbrukers kjønn, alder (i år), utdanning (høyskole/universitet), fulldyrket areal (dekar), hovedproduksjon (melk, korn, annet). Detaljert variabelbeskrivelse finnes i artikkelen.

Mange ville behandle en slik trinomisk analyse med responskategoriene 1, 2 og 3 som to logistiske regresjoner

$$\begin{aligned} &(1) \quad \text{logit}\, P(Y=3\,|\,\boldsymbol{x}) = \log \frac{P(Y=3\,|\,\boldsymbol{x})}{P(Y\le 2\,|\,\boldsymbol{x})} = \cdots \quad \text{og} \\ &(2) \quad \text{logit}\, P(Y\ge 2\,|\,\boldsymbol{x}) = \log \frac{P(Y\ge 2\,|\,\boldsymbol{x})}{P(Y=1\,|\,\boldsymbol{x})} = \cdots \end{aligned} \tag{8.4}$$

og kjøre dem separat. Det gir imidlertid to forskjellige sett med β_j-er som ikke representerer parallelle regresjonslinjer. De to løsningene ville ikke være kompatible med hverandre for alle verdier av $\boldsymbol{x}$.

Kjøring av eksempel:

```
PROC LOGISTIC DATA=bonder;
  CLASS arbeid (DESC) hoyutd mann melk korn / REF=first PARAM = REF;
  MODEL arbeid = alder hoyutd mann dekar01 melk korn / RSQ LACKFIT;
  WEIGHT vekt;
RUN;
```

Utskrift:

The LOGISTIC Procedure

Model Information

Data Set	WORK.BONDER
Response Variable	arbeid
Number of Response Levels	3
Weight Variable	vekt
Model	cumulative logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1552
Number of Observations Used	1534
Sum of Weights Read	1551.994
Sum of Weights Used	1531.307

Response Profile

Ordered Value	arbeid	Total Frequency	Total Weight
1	3	590	620.78815
2	2	298	311.18516
3	1	646	599.33389

Probabilities modeled are cumulated over the lower Ordered Values.

NOTE: 18 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information

Class	Value	Design Variables
hoyutd	0	0
	1	1
mann	0	0
	1	1
melk	0	0
	1	1
korn	0	0
	1	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
93.0889	6	<.0001

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	3241.164	2641.955
SC	3251.835	2684.641
-2 Log L	3237.164	2625.955

R-Square 0.3286 Max-rescaled R-Square 0.3740

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	611.2081	6	<.0001
Score	519.2036	6	<.0001
Wald	435.7861	6	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
alder	1	185.9236	<.0001
hoyutd	1	39.6843	<.0001
mann	1	18.5974	<.0001
dekar01	1	83.3155	<.0001
melk	1	153.6610	<.0001
korn	1	27.9399	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	3.3321	0.3065	118.1829	<.0001
Intercept	2	1	4.5025	0.3172	201.4290	<.0001
alder		1	-0.0736	0.00540	185.9236	<.0001
hoyutd	1	1	1.0563	0.1677	39.6843	<.0001
mann	1	1	0.7135	0.1655	18.5974	<.0001
dekar01		1	-0.00431	0.000472	83.3155	<.0001
melk	1	1	-1.5953	0.1287	153.6610	<.0001
korn	1	1	0.8350	0.1580	27.9399	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
alder		0.929	0.919	0.939
hoyutd	1 vs 0	2.876	2.070	3.995
mann	1 vs 0	2.041	1.476	2.823
dekar01		0.996	0.995	0.997
melk	1 vs 0	0.203	0.158	0.261
korn	1 vs 0	2.305	1.691	3.141

Teksten `Probabilities modeled are cumulated over the lower Ordered Values` betyr at modellen estimert egentlig er den omvendte av den som er beskrevet ovenfor, eller med andre ord modellen for

$$(1) \qquad \text{logit}\, P(Y=3) = \log \frac{P(Y=3)}{P(Y \le 2)} = \cdots \text{ og}$$

$$(2) \qquad \text{logit}\, P(Y \ge 2) = \log \frac{P(Y \ge 2)}{P(Y=1)} = \cdots$$

Dette er egentlig samme modellen med motsatt fortegn og forklaringen på at `Intercept 3 = 3.3321 < Intercept 2 = 4.5025.`

En annen type logiter som kan brukes for ordinale kategorier er såkalte "tilstøtende kategori logiter" (adjacent categories logits). De er definert ved

$$\log(\frac{\pi_{j+1}}{\pi_j}) = \alpha_j + \beta_j x_j, \quad j = 1,\ldots,J-1 \qquad (8.5)$$

Den multinomiske for nominale variable som logitmodellen som `LINK = glogit` spesifiserer er av formen

$$\log(\frac{\pi_j}{\pi_J}) = \alpha_j + \beta_j x, \quad j = 1,\ldots,J-1 \qquad (8.6)$$

Her er kategori *J* tatt som referansekategori, men den kan velges fritt. Merk at β_j i (8.5) kan estimeres ved å ta $\beta_{j+1} - \beta_j$ i modellen (8.6).

En tredje form for ordinal logit er *kontinuasjons-rate* logit. Den er av formen $\log(P(Y = j)/P(Y \le j-1)), j = 1,\ldots,J-1$ eller om en vil:

$$\log(\frac{\pi_2}{\pi_1}),\log(\frac{\pi_3}{\pi_1+\pi_2}),\ldots,\log(\frac{\pi_J}{\pi_1+\pi_2+\cdots+\pi_{J-1}}) \qquad (8.7)$$

Et alternativ er å definere denne logiten fra "andre enden", $\log(P(Y \ge j+1)/P(Y = j))$, $j = 1,\ldots,J-1$ eller

$$\log(\frac{\pi_2+\cdots+\pi_J}{\pi_1}),\log(\frac{\pi_3+\cdots+\pi_J}{\pi_2}),\ldots,\log(\frac{\pi_J}{\pi_{J-1}}) \qquad (8.8)$$

De to modellene (8.7) og (8.8) er ikke ekvivalente. Denne modellen kan også estimeres med PROC LOCISTIC, men det forutsetter at en først manipulerer data ved å aggregere responskategorier svarende til summene $\pi_1+\pi_2+\cdots+\pi_k,(k<J)$ i (8.7) eller $\pi_k+\cdots+\pi_J,(k>1)$ i (8.8).

9. Dispersjon

Metodene som er beskrevet for logistisk regresjon forutsetter at alle responser skjer uavhengig av hverandre. Er dette tilfellet vil

$$E(Y_i \mid \mathbf{x}_i) = \pi(\mathbf{x}_i), Var(Y_i \mid \mathbf{x}_i) = \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \text{ og } Cov(Y_i, Y_j \mid \mathbf{x}_i, \mathbf{x}_j) = 0 \quad (9.1)$$

for alle observasjoner *i* og alle par av observasjoner *i,j*. Hvis (9.1) holder, *g* er en gruppe observasjoner med samme verdi på $\mathbf{x}_i$, n_g er antall observasjoner i gruppen, π_g er deres felles verdi for $\pi(\mathbf{x}_i)$ og Z_g er antall "positive" begivenheter i gruppen, vil

$$EZ_g = n_g\pi_g \text{ og } Var(Z_g) = n_g\pi_g(1-\pi_g)\,. \qquad (9.2)$$

Hvis kovariansene (og tilsvarende korrelasjoner) i (9.1) ikke holder vil variansen i (9.1) heller ikke holde og vi får det som kalle over- eller underdispersjon. Positive korrelasjoner mellom svarene gir overdispersjon som betyr høyere varians enn i (9.1) mens negative korrelasjoner gir underdispersjon og lavere varians. Overdispersjon er det mest vanlige. Tilsvarende gjelder for så vidt også selv om $\mathbf{x}_i$ (og derved $\pi(\mathbf{x}_i)$) ikke er den samme for alle i gruppen, men det er enklest å illustrere situasjonen hvor de er like.

Overdispersjon kan oppstå på en rekke forskjellige måter. En av de mest vanlige er klyngedannelser i populasjonen. Familier og husholdninger er typiske eksempler på slike naturlig forekommende klyngedannelser i populasjonen. Hvis det er større tendens til at medlemmene i samme familie eller husholdning har samme verdi for Y enn det vi kan forvente fra like mange tilfeldig valgte individer fra hele befolkningen, vil det oppstå det som kalles *intraklyngekorrelasjon*. Dette vil generere overdispersjon. Hvis det for eksempel er slik at ektefeller er mer like hverandre med hensyn på utdanning, inntekt eller sysselsettingsstatus enn hva vi kan forvente av to tilfeldig valgte individer, vil det oppstå intraklyngekorrelasjon og overdispersjon med hensyn til disse variablene. Intervjuereffekter kan også generere overdispersjon. Hvis noen intervjuere legger for dagen holdninger eller forventninger som kan påvirke respondenter til å svare i en bestemt retning vil det kunne genereres intraklyngekorrelasjon mellom svarene som avgis til samme intervjuer. I eksempelet med lastebiler vil intraklyngekorrelasjon og overdispersjon kunne tenkes å oppstå dersom et firma får trukket flere biler og tilbøyeligheten til å svare er kritisk avhengig av firmaets holdning til undersøkelsen.

Vi kan tenke oss at vi har k grupper som vi for enkelhets skyld kan anta er like store med n observasjoner i hver. Hele utvalget er da på $m = nk$ observasjoner. Totalt antall "begivenheter" som observeres i de k gruppene blir da $Z = Z_1 + Z_2 + \cdots + Z_k$.

Det kan vises at ubetinget forventning og varians til Z er

$$E(Z) = m\bar{\pi} \text{ som er upåvirket.}$$
$$Var(Z) = (1 + (n-1)\rho)m\bar{\pi}(1-\bar{\pi}) = \sigma^2 m\bar{\pi}(1-\bar{\pi}) \qquad (9.3)$$

$\sigma^2 = 1 + (n\text{-}1)\rho$ blir *overdispersjonsparameteren*. ρ kan tolkes som intraklyngekorrelasjonen. Den må ligge i intervallet $(-1/(n\text{-}1), 1]$. Formelen viser at jo større klynger, jo større blir overdispersjonen for en gitt intraklyngekorrelasjon.
`MODEL` inneholder to opsjoner som kan brukes til å beregne over/under dispersjonen, `AGGREGATE=` og `SCALE=keyword`. `AGGREGATE` spesifiserer i prinsippet de variablene som definerer gruppene som genererer intraklyngekorrelasjon. Teknisk kan hvilken variabel i datasettet brukes til å definere slike grupper. `SCALE` velger en metode for å estimere overdispersjonen (σ^2). Det er fire opsjoner, `PEARSON, DEVIANCE, WILLIAMS, NONE og konstant`. Den siste brukes hvis overdispersjonen er kjent, og når er den det? `WILLIAMS` kan bare brukes ved *event/trial* syntaks. `PEARSON og DEVIANCE` er basert på at Residual Sum of Square eller Devians blir beregnet innen hver av gruppene og så summert. Når disse målene divideres med sine frihetsgrader (antall grupper k ganger (antall responskategorier –1) minus antall parametre i modellen), estimeres σ^2. Hvis $\sigma^2 > 1$ blir alle varianser og kji-kvadratobservatorer tilsvarende mye for store. For detaljer viser jeg til PROC LOGISTIC manualen side 62-65 og McCullagh & Nelder (1989) side 124-128.

Ved bruken av de metodene er det flere forhold som kan skape problemer. Det er viktig at en "korrekt" modell er etablert. Hvis det mangler variable i modellen som burde ha vært med, kan det komme til uttrykk i en for høyt estimert dispersjon.

For eksempelets skyld vises resultatet av en kjøring der variablene som inngår i regresjonen selv brukes som grupperingsvariable.

```
MODEL deltok = alder kl3 nasj3 regi2 / AGGREGATE=(alder kl3 nasj3 regi2)
                                                    SCALE=PEARSON;
```

`alder kl3 nasj3 regi2` definerer 30 x 4 x 2 x 2=480 grupper. 313 av dem inneholder biler. Det er 7 parametre og 313 – 7 = 306 som blir antallet frihetsgrader for dispersjonsparameteren. Utskriften nedenfor hvor noen av resultatene uten `SCALE` er redigert inn, viser at alle Kji-kvadratstørrelser er dividert ned med den estimerte σ^2, 1,07707. Alle standardfeil er multiplisert opp med $\hat{\sigma} = \sqrt{1,07707} = 1,0378$. Parameterestimatene ellers er uforandret.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	306	333.9072	1.0912	0.1309
Pearson	306	329.5826	1.0771	0.1694

Number of unique profiles: 313

NOTE: The covariance matrix has been multiplied by the heterogeneity factor (Pearson Chi-Square / DF) 1.07707.

Model Fit Statistics

	Nå		Før	
Criterion	Intercept Only	Intercept and Covariates	Intercept Only	Intercept and Covariates
AIC	2149.741	2086.321	2315.261	2246.029
SC	2155.839	2129.007	2321.359	2288.715
-2 Log L	2147.741	2072.321	2313.261	2232.029

Testing Global Null Hypothesis: BETA=0

	Nå	Før		
Test	Chi-Square	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	75.4198	81.2322	6	<.0001
Score	76.2324	82.1074	6	<.0001
Wald	73.1888	78.8293	6	<.0001

Type III Analysis of Effects

		Nå	Før	
Effect	DF	Wald Chi-Square	Wald Chi-Square	Pr > ChiSq
alder	1	37.7504	40.6597	<.0001
kl3	3	22.3658	24.0894	<.0001
nasj3	1	8.2723	24.0894	0.0040
regi2	1	13.4083	8.9099	0.0003

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.7403	0.1742	247.3334	<.0001
alder		1	-0.0568	0.00924	37.7504	<.0001
kl3	4	1	0.3987	0.1588	6.3057	0.0120
kl3	5	1	-0.2546	0.1724	2.1801	0.1398
kl3	6	1	0.5538	0.3198	2.9999	0.0833
nasj3	1	1	-0.3787	0.1317	8.2723	0.0040
regi2	1	1	-0.4668	0.1275	13.4083	0.0003

En annen måte å angripe de problemstillingene som dispersjon reiser, er ved hjelp av flernivåmodeller, med de hører ikke hjemme her.

10. Noen andre kommandoer og opsjoner i LOGISTIC

Det finnes flere kommandoer med opsjoner i PROC LOGISTIC enn det vi har vært inne på så langt. En fullstendig liste ser ut som følger.

BY *variable* Denne kommandoen tillater oss å gjøre logistisk regresjon gruppevis der gruppene er definert ved en eller flere kategoriske variable. Ingen opsjoner

CLASS *variable* Denne har vi brukt med noen opsjoner

CONTRAST Tillater testing av kontraster

FREQ *variable* Kan definere en antallsvariabel. Kan erstatte WEIGHT hvis vektene bare er hele tall. Ingen opsjoner

MODEL Alltid med. Noen nyttige opsjoner som vi ikke har brukt vil bli forklart nedenfor.

OUTPUT Tillater etablering av nytt datasett med alle observasjonene og estimerte og/eller kryssvaliderte sannsynligheter i tillegg til alle de gamle variablene. Noen nyttige opsjoner som vi ikke har brukt vil bli forklart nedenfor.

TEST Tillater testing av mer generelle hypoteser enn CONTRAST.

UNITS

WEIGHT

10.1. Noen opsjoner som ikke er beskrevet tidligere

For CONTRAST:

ESTIMATE = parm| exp | both ber om estimering og konfidensintervaller for kontraster eller exponensierte kontraster eller begge deler.

ALPHA=verdi Spesifiserer signifikansnivå for konfidensintervaller laget av ESTIMATE.

For MODEL:

CTABLE Lager en tabell over klassifiseringer av responser hvor estimerte sannsynligheter ”avrundes” til 0 eller 1 observasjoner med. Avrundingen bestemmes av en ”cutpoint” grense *z* slik at sannsynligheter større enn *z* avrundes til 1 og sannsynligheter mindre enn *z* avrundes til 0. Kan brukes for binære og ordinale responser (flere kategorier)

PPROB=verdi eller liste Setter *z* verdier for CTABLE. Hvis flere responskategorier brukes en liste. Se manual.

INFLUENCE Ber om identifisering av observasjoner som har særlig stor innflytelse på estimeringen og diagnostiske mål for deres innflytelse. Stor innflytelse er gjerne knyttet til ekstreme verdier av noen numeriske (kontinuerlige) forklaringsvariable med verdier på responsvariabelen som er atypiske for verdiene til forklaringsvariablene.

INCLUDE = *n* Brukes sammen med SELECTION. Krever at de *n* første variablene listet i MODELS alltid skal være med i modellene. Dette er for eksempel nyttig hvis noen stratifiseringsvariable tas med i modellen.

LINK Gir mulighet til å gjøre probit regresjon eller komplementær log-log regresjon i stedet for logit.

NOINT Undertrykker konstantleddet i regresjonen.

OFFSET=variabel Definerer en numerisk variabel som skal være med i regresjonen "som den er", dvs. med regresjonskoeffissient 1.

RSQUARE Ber om et mål for regresjonstilpassing som svarer til R^2 i vanlig lineær regresjon. Målet er basert på likelihoodfunksjonen. Se manualen. side 49.

For OUTPUT:

OUT=SAS dataset Obligatorisk når **OUTPUT** brukes. Spesifiserer datasettet som det skal skrives til ved bruk av **OUTPUT**. Kan være det samme som input datasettet. Inneholder alle variable som er med på input datasettet pluss dem som er bestilt ved opsjonene nedenfor.

XBETA Ber om at $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ skrives ut for hver observasjon.

PREDICTED=varnavn Ber om predikerte verdier (estimerte sannsynligheter) blir skrevet ut for hver observasjon. Kan forkortes til PRED, PROB eller bare P.

PREDPROBS=(nøkkelord) (nøkkelord) kan være (I), (C) , (X) eller kombinasjoner av disse som (I X) etc. I ber om det samme som PREDICTED, men gir sannsynligheter for begge/alle responskategorier. X ber om tilsvarende kryssvaliderte sannsynligheter. Se kapittel 10. I tillegg genereres to variable, _FROM_ og _INTO_. _FROM_ er lik den observerte verdien på responsvariabelen. _INTO_ er den responskategorien som har høyest sannsynlighet med de x-verdiene som gjelder for den aktuelle observasjonen.

STDXBETA Estimert standardfeil til XBETA.

LOWER=varnavn Ber om nedre grense for konfidensintervall til sannsynligheten estimert i PREDICTED.

UPPER=varnavn Ber om øvre grense for konfidensintervall til sannsynligheten estimert i PREDICTED.

11. Kryssvalidering

Dette er foreløpig og overfladisk.

Når målet med estimeringen er prediksjon, er det viktig å kunne validere prediksjonen. I SSB ville det kunne f.eks. tenkes at vi ville bruke registervariable som forklaringsvariable i en modell og en variabel som bare er kjent fra utvalg som respons. Så ville vi kunne prøve å predikere variabelverdiene for dem som ikke var med i utvalgsundersøkelsen. Mye av arbeidet i Folke- og boligtellingen foregikk etter slike linjer.

En slik prediksjon må valideres. Å sammenligne de estimerte sannsynlighetene med de samme data som vi har brukt til å estimere sannsynlighetene med, slik vi gjør ved testing av modellens tilpassing, vil gi et for godt bilde av modellens prediksjonsevne. Validering må skje mot andre data. Den klassiske metoden, i mangel av noe bedre, har vært å dele et utvalg tilfeldig i to deler, en del som brukes til å estimere modellen med og en som brukes til å validere prediksjonsevnen på. De to

delutvalgene er enten like store eller estimeringsdelen er noe større. Så kan man etterpå estimere med hele utvalget for å få bedre estimater.

En metode som er bedre, men som også krever mer er *kryssvalidering*. Kryssvalidering bruker hele utvalget. Ideen er å ta ut en observasjon, estimere modellen på resten av utvalget og "predikere" den observasjonen som var utelatt. Dette gjøres så for alle observasjoner i datasettet. Det lages så et samlemål på prediksjonsevnen.

Det er klart at å gjennomføre dette helt gjennom på et større datasett vil kunne kreve enorm datakraft. For det lastebildatasettet med 3288 observasjoner som har vært gjennomgangseksempelet i dette notatet, vil dette si at det måtte lages en loop som gjennomførte estimeringen 3288 ganger.

I det følgende vil det bli brukt visse forkortelser for å henvise til referanser:

AL: Anne Sofie **A**brahamsen og Knut Ivar **L**åstad: SAS/INSIGHT. Interne dokumenter 2000/1.

For oppstart av SAS INSIGHT og innhenting av datasett henvises til denne.

Alle datasett er tilgjengelige på Q:\metodekurs\SM07\SASdata. Kopier dem og lagre dem på ditt eget område.

12. Vedlegg: Utskrifter fra de første kjøringene

12.1. Eksempel 1. Hjerteinfarkt

12.1.1. Kjøring med SAS INSIGHT

Hent opp datasettet i tabell 4 til SAS INSIGHT som beskrevet i AL `(Solutions->Analysis->Interactive Data Analysis->Merk Library, Dataset->Open)`. Datasettet ligger som `Q:\Metodekurs\SM07\hjerteinfarkt.sas`. Følg beskrivelsen i AL kapittel 9: **Analyze --> Fit(Y X)**. I dialogvinduet velger du *x* (`ALDER`) og *Y* (`CHD`). Fjern `AGRP` hvis denne står i **Group** ruten. Set `Antall` inn i **Weight** ruten.

1. Vanlig lineær regresjon: Klikk på **Method** knappen. I vinduet som åpner seg merker du av **Normal** under Response dist. og **Identity** (eller canonical) under Link function. Trykk OK. Klikk på Output. Behold de automatiske innstillingene i dette vinduet men klikk på Output variables. Merk av **Predicted** og **Linear Predictor** og klikk OK, OK og Apply. Resultatet kommer nå opp i et vindu og grafen på side 5 er hentet derfra. I datavinduet kommer det nå opp tre nye variable, P_CHD (Predicted), LP_CHD (Linear Predictor) og R_CHD. Med de spesifikasjonene som er valgt blir P_CHD = LP_CHD = $\hat{\alpha} + \hat{\beta}x$ og R_CHD = $Y - \hat{\alpha} - \hat{\beta}x$, residualet. Merk at noen av de predikerte verdiene blir negative.
2. Klikk vekk vinduet med resultatene. Hvis du brukte Apply under punkt 1 er du tilbake i **Fit(Y X)** vinduet. Velg **Method** og merk av **Binomial** under response. La alle andre innstillinger være som unde punkt 1. INSIGHT vil nå tilpasse regresjonen under sannsynlighetsmaksimering for binomisk responsmodell. Grafen i resultatvinduet vil dessverre ikke vise regresjonslinjen som nå ville blitt

noe slakere enn under punkt 1. Nye P_CHD, LP_CHD og R_CHD beregnes etter samme formel som i punkt 1.

3. Klikk vekk resultatvinduet igjen. Velg **Method** og logit link. Regresjonskurven (som ikke vises) skal ikke lenger være en rett linje. Nye P_CHD, LP_CHD og R_CHD beregnes i datasettet. Nå blir $\text{P\_CHD} = \hat{\pi}(x) = \exp(\hat{\alpha} + \hat{\beta}x)/(1 + \exp(\hat{\alpha} + \hat{\beta}x))$ mens $\text{LP\_CHD} = \text{logit}\,\hat{\pi}(x) = \hat{\alpha} + \hat{\beta}x$. R_CHD $= Y - \text{P\_CHD}$.

For å få ut grafikk med regresjonskurven som ikke kom med i punkt 2 og 3 må vi kjøre PROC LOGISTIC og PROC GPLOT.

12.1.2. Kjøring av ugrupperte data med PROC LOGISTIC

Program for kjøring av hjerteinfarktdata

```
LIBNAME PLOTLIB 'H:\Kurs\SM07\Grafikk';
/*
Enkel kjøring av modell for hjerteinfarkt etter alder.
Variabelen CHD er 1 hvis infarkt, 0 ellers.
DESCENDING nødvendig for å få modellering av CHD=1 i stedet for CHD=0;
*/
PROC LOGISTIC DATA=SM07.hjerteinfarkt DESCENDING;
  MODEL CHD = alder;          /* Bruker "single-trial" syntaks. */
  WEIGHT antall;
  OUTPUT OUT=SM07.hjinfarkt_ut L=nedre P=pihat U=ovre XBETA=xbeta;
/*
Generering av plott i figur 3.
*/
PROC GPLOT DATA=SM07.hjinfarkt_ut GOUT=PLOTLIB.hjinfarkt UNIFORM;
  SYMBOL1 INTERPOL=none C=black VALUE=none POINTLABEL=("#antall" J=c
POSITION=middle);
  SYMBOL2 INTERPOL=join CI=red L=20 ;
  SYMBOL3 INTERPOL=join CI=blue L=1 W=2 ;
  PLOT CHD*alder=1 nedre*alder=2 ovre*alder=2 pihat*alder=3 /OVERLAY;
RUN;
```

Utskrift fra LOGISTIC. **Beskrivelse av noen sammenhenger er innredigert.**

```
         Model Fit Statistics

                                          Intercept
               Intercept                    and
Criterion        Only                    Covariates

AIC             138.663 = -2 log L + 2·1     111.353 = -2 log L + 2·2
SC              140.837                      115.702
-2 Log L        136.663                      107.353

        Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        29.3099        1         <.0001
Score                   26.3989        1         <.0001
Wald                    21.2541        1         <.0001

             Analysis of Maximum Likelihood Estimates

                                 Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
```

```
Intercept     1   ( -5.3095   /  1.1337 )² =  21.9350        <.0001
Alder         1   (  0.1109   /  0.0241 )² =  21.2541        <.0001

           Odds Ratio Estimates

             Point                    95% Wald
Effect    Estimate                Confidence Limits

Alder     1.117          1.066                 1.171 = exp(0.1109+1.96·0.0241)
     = exp(0.1109)   = exp(0.1109-1.96·0.0241)
```

12.1.3. Kjøring av grupperte data med PROC LOGISTIC

Program

```
/*
Kjøring med grupperte data. Bruker gjennomsnittsalder i gruppen som
forklaring.
*/
PROC LOGISTIC DATA=SM07.infarkt_gr DESCENDING;
  MODEL Ja/N = snitt;              /* Bruker "Event/trial" syntaks. */
  OUTPUT OUT=SM07.infarkt_gr_ut L=nedre P=pihat U=ovre XBETA=xbeta;
RUN;

PROC GPLOT DATA=SM07.infarkt_gr_ut GOUT=PLOTLIB.infarkt_gr UNIFORM;
  SYMBOL1 INTERPOL=none C=black VALUE=none POINTLABEL=("#n" J=c
POSITION=middle);
  SYMBOL2 INTERPOL=join CI=red L=20 ;
  SYMBOL3 INTERPOL=join CI=blue L=1 W=2 ;
  PLOT andel_ja*snitt=1 nedre*snitt=2 ovre*snitt=2 pihat*snitt=3 /OVERLAY;
RUN;
```

Utskrift. For illustrasjon av sammenhenger, se utskrift i 4.1.2.

```
The LOGISTIC Procedure

                    Model Information

Data Set                          SM07.INFARKT_GR
Response Variable (Events)        Ja
Response Variable (Trials)        N
Number of Observations            8
Model                             binary logit
Optimization Technique            Fisher's scoring

                Response Profile

   Ordered      Binary           Total
     Value      Outcome      Frequency

         1      Event               43
         2      Nonevent            57

              Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	138.663	112.308
SC	141.268	117.518
-2 Log L	136.663	108.308

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.3552	1	<.0001
Score	25.7709	1	<.0001
Wald	21.0346	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.2869	1.1343	21.7236	<.0001
Snitt	1	0.1094	0.0239	21.0346	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Snitt	1.116	1.065	1.16

13. Referanser

Agresti, A. (1990): *Categorical Data Analysis.* Wiley.

---------- (1996): *An Introduction to categorical Data Analysis*. Wiley.

Cox, D.R. og Snell, E.J. (1989). *Analysis of Binary Data.* 2. ed. Chapman & Hall.

Hosmer, D.W. og Lemeshow, S. (2000): *Applied logistic regression.* Wiley.

McCullagh, P. og Nelder, J.A. (1989): *Generalized Linear Models.* 2. ed. Chapman &Hall Monographs on Statistics and Applied probability no. 37.

Zelterman, D. (1999): *Models for Discrete Data*. Oxford Science Publications.

De sist utgitte publikasjonene i serien Notater

2006/27 J. Heldal og A. Rusti: Om samordning av utvalg vedbruk av PRN-tall. 29s.

2006/28 C. Nordseth og Ø. Sivertstøl: FD - Trygd. Dokumentasjonsrapport. Fødsels- og sykepenger, 1992-2003. 134s.

2006/29 A. Linderud: Verdipapirstatistikk. Dokumentasjonsnotat. 54s.

2006/30 V.V. Holst Bloch, H. Høye, M. Steinnes og J.K Undelstveit: Kartbasert rapportering i KOSTRA - en mulighetsstudie. 50s.

2006/31 E. Høydal: Monitor for sekundærflytting. En deskriptiv analyse av sekundærflyttinger blant flyktninger bosatt i Norge i 1995-2004. 67s.

2006/32 E.Cometa Rauan: Undersøking om foreldrebetaling i barnehagar, januar 2006. 46s.

2006/33 T. Skarðhamar: Kriminalitet gjennom ungdomstiden blant nordmenn og ikke-vestilige innvandrere. En analyse av fødselskullet 1977. 36s.

2006/34 N. Hagesæther og L-C. Zhang: Om arbeidsledighet i AKU og Arena. 19s.

2006/35 T. Hægeland, Lars J. Kirkebøen og Oddbjørn Raaum: Skoleresultater 2005. En kartlegging av karakterer fra grunnskoler og videregående skoler i Norge. 83s.

2006/36 S. Skaare: Undersøkelse om «Utbrenthet i enkelte yrker» 2005. Dokumentasjonsrapport. 68s.

2006/37 O.F. Vaage: Barn og unges idrettsdeltakelse og foreldres inntekt. Analyse med data fra Levekårsundersøkelsen 2004. 31s.

2006/38 A.Vedø og L. Solheim: En praktisk innføring i utvalgsplanlegging. 40s.

2006/39 H.C. Hougen: Samordnet levekårsundersøkelse 2005 - tverrsnittsundersøkelsen. Dokumentasjonsrapport. 156s.

2006/40 T. Nøtnæs, S. Bytingsvik og B. Hole: Resultater fra brukertesting av ssb.no. 34s.

2006/41 KOSTRA. Arbeidsgrupperapporter 2006. 169s.

2006/42 T. Gulbrandsen: Levekårsundersøkelse blant studenter. Dokumentasjonsrapport. 66s.

2006/43 A-G. Jørstad: Overvåkingssystemet for bedrifter i Bof. 19s.

2006/44 M. høstmark og B.O. Lagerstrøm: Undersøkelse om Arbeidsmiljø: Destruktiv atferd i arbeidslivet. Dokumentasjonsrapport. 43s.

2006/45 T.K. Schjerven og K.Å. Wass: Faglig modell og rammeverk i StatRes. 67s.

2006/47 K. Henriksen: Utvalgsplan til konsumprisindeksens nye matvareindeks - Basert på strekkodedata. 23s.

2006/48 A.B. Thorud, D. Rafat, S. Ferstad og E. Vinju: Tverrgående revisjon i KOSTRA - Bedring av påliteligheten i nøkkeltallene. 65s.

2006/49 T. Granseth: Grensehandel. En analyse av kvaliteten av data. 48s.

2006/50 E. Engelien, H. Høie og M. Steinnes: Bygging i strandsona. Metode og resultater. 18s.