HSF
HØGSKULEN I
SOGN OG FJORDANE

NOTAT

Statistisk analyse med SPSS

Kristin Linnerud, Ove Oklevik,
Harald Slettvoll

N NR 13/2004 AVDELING FOR ØKONOMI OG SPRÅK

HSF
HØGSKULEN
I SOGN OG FJORDANE

Postboks 133, 6851 SOGNDAL telefon 57676000 telefaks 57676100

NOTAT

TITTEL	NOTATNR.	DATO
Statistisk analyse med SPSS	13/04	26/11-2004
PROSJEKTTITTEL	**TILGJENGE**	**TAL SIDER**
Internt finansiert/eigentid		88

FORFATTAR
Kristin Linnerud
Ove Oklevik
Harald Slettvoll

PROSJEKTLEIAR/-ANSVARLEG
Kristin Linnerud

OPPDRAGSGJEVAR

EMNEORD
Statistisk analyse, varians-, regresjons-, faktor- og diskriminantanalyse

SAMANDRAG

Dette notatet har sitt utspring i forelesninger og undervisning for 3.års studenter i økonomi og administrasjon ved høgskolen i Sogn og Fjordane. Notatet er særlig lagt opp mot undervisningen i SPSS i de to kursene "OR 685 Marknadsanalyse og merkevarestrategi" og "BD 616 Økonomistyring og analyse med programvare".

SUMMARY

The primary objective of the paper is to give an introduction to how statistical anlysis can be conducted using the statistical software program SPSS. The paper started as a set of lecture notes for the two undergraduate courses "OR 685 Marketing research and branding strategies" and "BD 616 Applied Financial Analysis.

PRIS	ISSN	ANSVARLEG SIGNATUR
	0806-1696	

Forord

Dette notatet har sitt utspring i forelesninger og undervisning for 3.års studenter i økonomi og administrasjon ved høgskolen i Sogn og Fjordane. Notatet er særlig lagt opp mot undervisningen i SPSS i de to kursene ”OR 685 Marknadsanalyse og merkevarestrategi ” og ”BD 616 Økonomistyring og analyse med programvare”. Notatet egner seg imidlertid godt også til selvstudium i SPSS. Da tilgjengelig litteratur har vært spredt rundt på forskjellige læreverk og til dels er skrevet på engelsk, så vi behovet for en dokumentasjon som er bedre tilpasset undervisningen ved vår høgskole.

Sogndal 26/11 2004.

_____	_____	_____
Kristin Linnerud	Ove Oklevik	Harald Slettvoll

Innhold

1 Før du starter opp

1.1 Arbeidsmåte

For å foreta en statistisk analyse i SPSS gjør du følgende:

1. Legg inn dataene inn i vinduet "data editor"
2. Velg en prosedyre fra menyene
3. Velg variabler for analysen
4. Eksaminer resultatene

Du kan åpne en tidligere lagret datafil, lese et regneark, tekstfil eller database, eller taste inn dataene direkte i delvinduet ”Data Editor”.

Du velger deretter en prosedyre fra menyene for å lage tabeller, grafer eller produsere statistiske parametere.

Det dukker opp en dialogboks for hver prosedyre der du blir bedt om å definere hvilke variabler som skal analyseres. Her må du også presisere din bestilling av analyse mer nøyaktig.

Når du er fornøyd trykker du OK, og resultatene presenteres i et eget delvindu; ”SPSS Viewer”.

1.2 Hjelp til selvhjelp

SPSS har utviklet et system for hjelp til å lære seg systemet og løse problemer underveis:

Hjelpe-meny: Hvert vindu har en hjelpe meny i menylisten øverst. Under ”Topics” kan du søke i et stikkordsregister for å finne det emne du søker hjelp om. Brukerveiledning for alle deler av programmet finnes under ”Tutorial”. Under ”Case studies” finner du framgangsmåten for å løse konkrete statistiske problemstillinger. En ”Statistic Coach” fungerer som veiviser mht hvilke prosedyrer som egner seg for ulike datatyper og problemstillinger og hvordan du tolker resultatene.

Hjelpe-knapper i dialogboksene: De fleste dialogbokser har hjelpe knapper som bringer deg direkte til det emne du har spørsmål om og relaterte problemstillinger.

Pivot-tabeller: Klikk med høyre musetast på en aktivert pivot tabell i ”Viewer” og velg ”What’s This?” fra menyen som dukker opp.

1.3 Målenivå

Variable kan måles på ulikt nivå. Dette vil bestemme hvilke statistiske prosedyrer som kan brukes. Vi har følgende målenivå:

Nominelt: Hver dataverdi for en variabel er bare en gruppeidentifisering som for eksempel yrkeskategori, bosted eller lignende. Dataene kan kodes (by:2 og distriktet:1). Det er likevel ingen skalaretning på dataene, og vi kan heller ikke snakke om gjennomsnitt eller standardavvik for disse variablene. Et unntak er dikotome variable (ja/nei). For eksempel: ” Stemmer du Arbeiderpartiet?” Valget er: ja (1) og nei (0). Her kan vi tolke gjennomsnittet som andelen som stemmer Arbeiderpartiet. Slike variable kan brukes i statistisk sammenheng som kontinuerlige variable.

Ordinalt: Dette er variable som kan rangeres men som ikke følger en bestemt skala. Eksempler er svaralternativene: 1: helt enig, 2: litt enig, 3: litt uenig og 4: helt uenig. Vi kan ikke si noe om avstanden på 1 og 2, men vi vet skalaretningen. Strengt tatt kan man ikke regne gjennomsnitt av ordinale variable. Dette blir likevel ofte gjort. Resultatene kan bli bra viss vi har et stort antall kategorier. Antall kategorier bør generelt settes lik et oddetall. Vi bør inkludere et valg som heter ”vet ikke” evt. ”manglende svar”. Alternativt kan man lage en skala fra ”helt uenig” til ”helt enig”, og be om at man setter et kryss på linjen (Likerts skala). Avmerkingene kan leses vha en skanner og omgjøres til numeriske kontinuerlige data.

Kontinuerlig: Dette er numeriske data med lik avstand mellom hvert punkt på skalaen. De kontinuerlige data som har et nullpunkt (f. eks alder) kalles metriske variabler. Du kan utføre de fleste typer statistiske tester på kontinuerlige data såfremt gitte forutsetninger om for eksempel normalfordelte data etc. er tilfredsstilt.

1.4 Filer

I gjennomgangen av ulike problemstillinger vil jeg bruke en rekke datafiler. Disse blir gjort tilgjengelig for studentene via Classfronter. Oversikt over hvilke filer som er brukt er gitt til slutt i dette notatet.

1.5 Versjon av SPSS

Vi har brukt SPSS versjon 12.0 ved utarbeiding av dette notatet.

2 Legge inn og redigere data[1]

Klikk på ikonet for SPSS, evt. finn programmet under ”Start” og ”Alle programmer”. Får du opp en boks med en rekke spørsmål, trykker du ”cancel”. Du står da i delvinduet "Data Editor".

2.1 Legge inn data

Når du legger inn data, må du først definere variablene i delvinduet "Variable View". Deretter legger du inn dataene i delvinduet "Data View".

For å ***legge inn data manuelt*** trykker vi:

File
New
Data

Vi får et blankt dataregistreringsark. SPSS datafiler er organisert med ett kasus for hver rekke og variablene i kolonner. Et kasus kan være et ferdig utfylt spørreskjema for en respondent. Svaret på hvert spørsmål i spørreskjemaet er gitt i kolonnene.

Du skal nå legge inn:

Idnr:	kjønn:	høyde:
10000	mann	1,87
20000	kvinne	1,63
30000	kvinne	1,50
40000	mann	1,75

Noen kjøreregler for innlegging er:

- I alle datasett bør det være idnummer eller lignende som entydig angir hvilket kasus dette er.
- Dataene bør helst kodes som numerisk informasjon (f.eks : mann=1 og kvinne=0)
- Dataene bør være mest mulig på grunnlagsform. Det er for eksempel bedre med fødselsdato enn alder eller aldersgruppe.

Vi går inn i delvinduet "Variable View". Vi skal nå definere egenskapene til variablene idnummer, kjønn og høyde før vi legger inn data.

Idnr: Vi plasserer markøren i første linje under "Name". Vi har 8 tegn til å definere variablenavn og det første tegnet må være en bokstav. Vi

[1] Dette avsnittet er i stor grad basert på/inspirert av kursmaterialet ved et SPSS kurs 2001 i regi av Capture Data as.

skriver: "Idnummer". Vi trykker tabulator og en rekke av de andre cellene blir forhåndsutfylt. Under "Type" lar vi det stå "Numeric". Under "Width" reduserer vi tallet til 5. Under "Decimals" setter vi tallet lik 0. Vi kan angi en lengre forklarende tekst under "Labels". Maks antall tegn er 80, men vær oppmerksom på at denne teksten blir brukt i grafer og lignende og her kan grensen være rundt 20 tegn. De andre feltene lar vi stå uendret.

Kjønn: Vi legger deretter inn kjønn som variabelnavn på neste linje. Denne variabelen legger vi også inn som et tall; 0 eller 1. Antall tegn blir derfor 1. Legg merke til at du må redusere tall desimaler til 0 før du får lov å sette antall tegn lik 1. Under ”Values” legger du så inn betydningen av kodene 0 og 1 for kjønn. Trykk på ”…” i cellen. I boksen som kommer opp legger du inn 1 under ”Value” og mann under ”Value Label” og trykker ”Add”. Deretter gjør du det samme for kjønn = 0. Husk igjen å trykk ”Add”! Du har nå kodet informasjonen som ligger i 0 og 1.

Høyde: Du legger dette inn på samme måte som for idnummer. Antall siffer kan her settes lik 4. Da har du tatt høyde for at også komma og de to desimalene krever hvert sitt tegn.

Vi går nå inn i delvinduet "Data View" og fyller inn dataene. Under menyvalget "View" kan vi plassere markøren på "Value Labels" og venstreklikke for å skru denne på. Når du har haket av ”Value Labels, kan vi fylle inn data for ”Kjønn” ved å skrive "mann" eller "kvinne", 1 eller 2 eller bruke rullgardinen. Forsøk å endre innstillingen og se hva som skjer!

Dataene ***lagres*** ved å trykke:

File
Save as
Filnavn: ”HSFtest”

Vi kunne alternativt ha ***hentet dataene fra en fil***. Vi går da opp i menyen og trykker:

File
Open
Data

Oppgave: hent inn filen **"Gujarati tab 6 3. sav".**

Datafilene vi bruker i SPSS trenger ikke være på SPSS format; dvs. med endelsen .sav. Forsøk å hente inn Excel-filen **”Tabell avling og gjodsel.xls”**.

2.2 Retting og redigering av innlagte data

Du kan fritt redigere dine registeringer i ”Data Editor”. Du kan ***endre på egenskapene*** i delvinduet "Variabel View" og ***endre dataene*** i delvinduet "Data View". Du kan ***endre rekkefølgen*** på variablene i delvinduet "Data View" ved først å plassere markøren på overskriften til variabelen og markere kolonnen med venstre musetast. Deretter trykker du venstre musetast en gang til, samtidig som du drar kolonnen bort til ønsket plassering.

Det er viktig å ***lese korrektur*** på innlagte data. Kanskje har du lagt inn data som er utenfor parameterområdet; for eksempel 3 under kjønn. Eller du har lagt inn 1,87 i stedet for 1,78 under høyde. Den første feilen er mer graverende enn den siste. En slik vill verdi kan ødelegge resultatet av en del prosedyrer som for eksempel regresjonsanalyse.

Hent fram filen: ”**HSFtest.sav**” som du laget i stad. Legg så inn kjønn lik 2 på et av observasjonene. En slik opplagt feil verdi kan vi finne ved å ved å kjøre prosedyren:

Analyze
- Deskriptiv statistics
  - Deskriptiv (kontinuerlige)
  - Frequencies (nominelle/ordinale variable)

Vi markerer ønsket variabel, ”piler” den over i høyre boks og trykker OK. Hvis den resulterende utskriften viser en maksverdi kjønn er større enn 1, har vi tastet feil.

Ved store datamengder kan det være greit å ***søke fram*** hvilket kasus som ga denne verdien. Dette gjør vi slik:

Vi markerer kolonnen "kjønn" i vinduet ”Data Editor”. Vi går så opp i menyen og velger:

Edit
- Find

I dialogboksen som dukker opp legger vi inn den observerte ekstremverdien på kjønn, og vi får opplyst det første kasus som har denne verdien. Verdien rettes så i ”Data Editor” og vi lagrer på nytt.

2.3 Rekode

Vi har som hovedregel at nye data bør legges inn på såkalt grunnlagsform. I konkrete analysesituasjoner kan det likevel være bruk for tallene på en annen form.

Vi skal i følgende eksempel bruke filen: **”Demo.sav”**[2].

Eksempel: Vi ønsker å dele kasusene inn i tre kategorier avhengig av antall medlemmer i husholdningen (eng: residents):

Gammel variabel:	Ny variabel:	
resident [1]	husholdt = 1	(kodes = single)
resident [2-4]	husholdt = 2	(kodes = liten husholdning)
resident [5 ->]	husholdt = 3	(kodes = storhusholdning)

Vi velger fra menyen:

Transform
Recode
Into Different Variable

Her velger vi variabelen med ”Reside”[3] (variabelen som skal omkodes) inn i boksen ”Numeric Variable”. Så legger vi navnet på den nye variabelen ”Husholdt” inn i boksen ”Output variable”, ”Name”. Vi klikker på ”Change” og ser at det nye variabelnavnet flyttes opp i boksen ”Numeric Variable; Output”.

Vi velger så ”Old and New Variables” for å angi hvilke verdier på den gamle variabelen som skal knyttes opp mot bestemte verdier i den nye variabelen. Den første omkodingen skjer ved at du velger 1 som verdi under ”Old Value” og 1 som verdi under ”New Value”. Trykk ”Add”. De to neste omkodingene skjer på samme måte, men merk at du krysser av for ”Range” for å angi intervallet [2-4] og ”All other Values” for å angi intervallet [5,->) under ”Old Values”.

Du kan til slutt navngi de tre husholdningstypene på samme måte som vi gjorde for variabelen ”Kjønn” i filen ”HSFtest.sav”,

I inputmatrisen har vi nå fått en ny variabel som heter ”Husholdt”. Denne kan vi bruke på vanlig måte i ulike analyser. Men husk: en del analyser krever kontinuerlige data!

2.4 Compute

Skal vi gjøre omkodinger som er av en slik form at den nye variabelen baseres på en matematisk formel eller funksjon av en eller flere definerte variable, så må vi bruke ”Compute” i stedet for ”Recode” i SPSS.

Hent inn filen: **”Banknor.sav”**.

[2] En oversikt over alle filer med kildereferanse og beskrivelse av innhold finner du bakerst i dette notatet.
[3] Det utvidede navnet til variabelen ”reside” er ”Number of people in houshold”. I tabellen med variabler til venstre, vil variabelens ”labell” være oppgitt før det korte variabelnavnet.

Vi ønsker for hver person i Banknor-filen å finne ut hvor mye personen har tjent i banken på den tiden han/hun har arbeidet der. En tilnærming kan være å regne ut formelen:

(Startlønn + Sluttlønn)/2*Ansettelsestid i måneder = Tjent Totalt

Vi velger fra menyen:

Transform
Compute

Under ”Target Variabel” skriver vi "Tjenttot".

Under ”Numeric expression” skriver vi:
”((slutlønn+startløn)/2)*(anstmnd/12)”.

Alternativt kunne vi skrevet det slik:
”MEAN(startløn,slutlønn)*(anstmnd/12)”.

Her har vi brukt standardfunksjonenen ”MEAN” fra ”Functions” til høyre i dialogboksen.

Velger vi OK, får vi ”Tjenttot” til slutt i inndatamatrisen.

Her finnes det mange muligheter. Trykk på F1 så får du hjelp til de ulike standardfunksjoner. For eksempel kan funksjonen ”CTIME.DAYS()” regne ut alderen til hver person i et spørreskjema ved å legge inn dato for spørreundersøkelsen - fødselsdato inne i parantesen. Det forutsetter selvsagt at disse to variabelene er opprettet i inputmatrisen.

Under ”Type & Labell” kan vi legge inn forklarende tekst til variabelen ”Tjenttot”; for eksempel: ”tjent totalt i banken”.

NB! Gjør du endringer i variablene som inngår i slike funksjoner etter at du har kjørt Compute funksjonen, så blir ikke den nye target variabelen endret! Da må du kjøre Compute en gang til.

2.5 Utvalgte kasus

Vi bruker ”Select Cases” når vi skal gjøre statistiske analyser av en spesiell delmengde av datamatrisen.

Hent inn datafilen: **Demosav.sav**.

Fra menyen velger vi:

Data
Select cases

I dialogboksen som dukker opp er standard innstilling "All Cases". Etterfølgende statistikk blir kjørt på alle kasusene. Velger vi imidlertid "If Condition is Satisfied" kan en trykke If-tasten og legge inn vilkår for hvilke kasus som skal være med i den statistikken som følger. Her skriver vi vanlige matematiske funksjoner som skal avgrense hvilke kasus som skal være med. Vanlige matematiske operatorer som ”AND” og ”OR” samt andre tegn som for eksempel >,< osv. gjelder. Bruk gjerne paranteser for å klargjøre spillereglene.

Velg videre fra menyen:

If Condition is Satisfied

Vi ønsker nå kun å analysere delutvalget som har inntektskategori 2. I If boksen legger du inn betingelsen ”Inccat = 2”[4] . Trykk ”Continue” og ”OK”. Deretter velger vi fra menyen:

Analyze
Descriptive statistics
Frequencies

Vi ønsker å vise deskriptive data for variabelen bilkategori. Vi piler inn variabelen ”Carcat” og trykker ”OK”. Vi får da:

Primary vehicle price category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Economy	667	27,9	27,9	27,9
	Standard	1721	72,1	72,1	100,0
	Total	2388	100,0	100,0	

I inputmatrisen blir det satt skråstrek over de kasusene som ikke tilfredsstiller betingelsen i ”Select Cases” valget. Dette skjer i nummereringslinjen til venstre. Vi får og en ny variabel i inputmatrisen som kalles ”Filter_$”. Vi får et varsel på nederste linje om ”Filter On”.

I utskriften ser vi at antall kasus nå er redusert til 2388. Skal vi klargjøre at vi har brukt et filter, kan vi i utskriftsfilen legge inn en ny overskrift. Klikk med venstre musetast på tabellen. Velg ”Insert” og så ”New Title” fra menyen. Skriv så ”Statistikk gjelder kun inntektsgruppe = 2” i det avmerkede feltet som dukker opp.

NB! Skal vi igjen over til å bruke hele datamatrisen, må vi gå aktivt inn i Menyen og velge:

Data
Select Cases

[4] Denne variabelen har ”Label” : Income category in Thousands.

Krysse av "All cases".

2.6 Splitte filer

Skal vi foreta samme statiske analyser for flere verdier av en bestemt variabel, bruker vi valget ”Split File”. For eksempel kan vi ønske å utføre samme analyse som over, men for alle inntektskategoriene. Vi velger fra menyen:

Data
    Split File

Her er standard innstilling "All Cases". Den bytter vi ut med "Organize Output by Groups". I boksen som dukker opp velger vi variabelen ”Inccat” under valget ”Groups Base don”. Vi velger deretter fra menyen:

Analyze
    Descriptive Statistics
        Descriptive

Vi velger å analysere de kontinuerlige dataene pris på hovedbil og antall mennesker i husholdningen. Nedenfor er vist utskriften for kun den første inntektskategorien:

Income category in thousands = Under $25

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean		Std.	Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error
Price of primary vehicle	1174	4,20	13,00	9,3776	,0588	2,01626	-,279	,143
Number of people in household	1174	1	9	2,49	,05	1,546	,342	,143
Valid N (listwise)	1174							

a. Income category in thousands = Under $25

Vi merker oss at ”Split File” ikke gir skråstreker over enkeltkasus slik som ”Select Cases”. Vi får ingen filtervariabel registrert i inputfilen. Vi får en påminnelse om at ”Split File” er valgt nederst til høyre i inputfilen
Til slutt slår vi av ”Split File” kommandoen fra menyen:

Data
    Split file

Her krysser vi av for "All cases".

NB! Vi kan alternativt få tilsvarende resultat gjennom å bruke ”Compare Means”. Prøv!

3 Tabeller, diagram og grafer

3.1 Frekvenser

Vi tar inn filen: **”Demo.sav”** og gjør følgende valg:

Analyze
Descriptive statistics
Frequencies

Standard utskrift

Vi velger vanligvis nominelle eller ordinale variabler. Vi kan og velge metriske variable, men vi kan fort få svært lange frekvenstabeller som resultat. Velger vi for eksempel ”Income Category in Thousand” uten andre avkrysninger får vi følgende tabell:

Income category in thousands

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under $25	1174	18,3	18,3	18,3
	$25 - $49	2388	37,3	37,3	55,7
	$50 - $74	1120	17,5	17,5	73,2
	$75+	1718	26,8	26,8	100,0
	Total	6400	100,0	100,0	

Vi får en utskrift med ”Value Labels” (beskrivelser), frekvensopptelling samt kumulativ prosent.

Statistiske parametere

Vi gjentar nå menyvalgene over, men velger denne gangen en kontinuerlig variabel som for eksempel inntekt i husholdningen. I dialogboksen fjerner vi krysset for ”Display Frequency Table**”** og krysser i stedet av for ”Statistics”. En ny dialogboks dukker opp, og vi gjør valg av ulike statistiske parametere som ”Mean”, ”Median”, ”Kurtosis”, ”Skewness”, ”Std. Error of Mean” etc.. Avhengig av våre valg, kan vi få et resultat som følger:

Statistics

Household income in thousands

N	Valid	6400
	Missing	0
Mean		69,4748
Std. Error of Mean		,98398
Median		45,0000
Std. Deviation		78,71856
Skewness		4,513
Std. Error of Skewness		,031
Kurtosis		33,877
Std. Error of Kurtosis		,061

Legg merke til at estimering av denne type statistiske parametere ofte krever data på kontinuerlig skala.

Når du står inne i dialogboksen ”Statistics”, kan du få forklaring på hva de ulike begrepene betyr. Plasser markøren på for eksempel ”Skewness” og høyreklikk. Alternativt kan du klikke direkte på hjelp i dialogboksen. Noen forklaringer følger:

MEAN: det aritmetiske gjennomsnittet. Her er gjennomsnittlig månedslønn 13.767,83.

STD ERR: Standardfeilen for en gjennomsnitts månedslønn. Dette målet sier noe om usikkerheten til gjennomsnittslønnen over. Matematisk uttrykt er det lik STD. DEV/Roten av antall kasus

MEDIAN: Dette er den midterste lønnen når vi sorterer månedslønnene fra lavest til høyest. Halvparten tjener mer enn denne lønnen og halvparten tjener mindre.

MODE: Den lønnen som er mest frekvent.

STD.DEV: Gir standardavviket (et spredningsmål) for de enkelte kasus sin spredning rundt gjennomsnittet.

VARIANCE: Er variansen til målingene. Varians = STD.DEV*STD.DEV. Det er ikke nødvendig å oppgi begge målene for spredning.

KURTOSIS: Er et mål som sier noe om "skjørtekantene" i en fordeling. Er verdien i kurtose i nærheten av null, så tyder det på at fordelingen er tilnærmet normalfordelt. Negative verdier indikerer at variabelen har mindre deler av fordelngen i halene enn tilsvarende normalfordeling. SE KURT er standardfeilen til KURTOSE, et spredningsmål som sier noe om usikkerheten i Kurtosemålet.

SKEWNESS: Er et mål som sier noe om fordelingen er symmetrisk. Normalfordelingen (og andre symmetriske fordelinger) har Skjevhet = 0. Skjevhetsmålet er slik at hvis det har positiv verdi har tilsvarende fordeling en opphopning i fordelingen til høyre haledel, mens en negativ verdi viser en opphopning i

venstre haledel. Skjevhet er et av de mål (sammen med Mean, Median og Kurtose) som brukes til å angi om en variabel er tilnærmet normalfordelt. SE SKEWNESS er standardfeilen til Skjevhet. Et spredningsmål for Skjevhet.

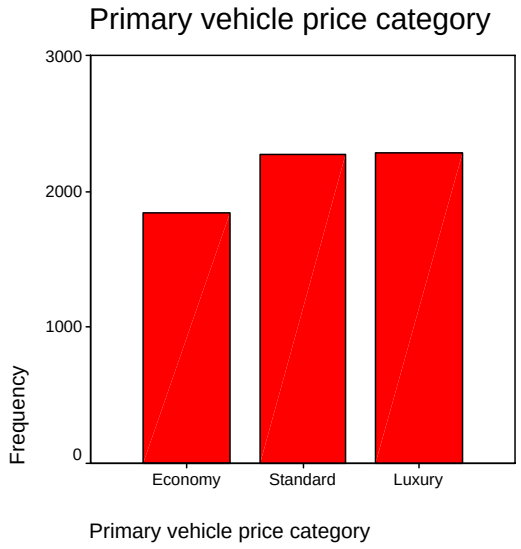
Tommelfingerregel: målene KURTOSIS og SKEWNESS bør begge ha en absoluttverdi <1,5. Men dette avhenger bl.a. av antallet. For observasjoner på 30-40 kan for eksempel et krav om absoluttverdi mindre eller lik 1 være rett.

RANGE Er forskjellen mellom maksimumsverdien og minimumsverdien.

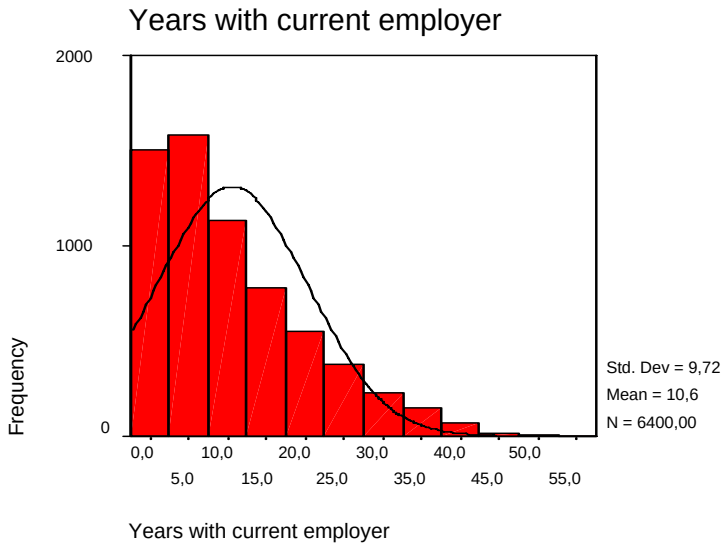
Vi kan og velge ”Percentile Statistics” i dialogboksen ”Statistics”. Her kan vi velge kvartiler som er de verdier i fordelingen som deler det rangerte materialet i 4 like store deler (jfr. Median). Vi kan og velge ”Cut Points for 10 equal points” og dermed få desentiler som altså er de verdier som deler materialet opp i 10 like grupper.

Grafikk

Inne i dialogboksen kan vi og krysse av for ”Charts”. Vi kan velge mellom ”None”, ”Bar Chart”, ”Pie Chart”, ”Histogram” med og uten normalkurve samt utforming av aksene ved ”Bar Chart”. Standard valg er ”None”. Velger vi ”Barchart” får vi stolpediagram. Nedenfor er vist stolpediagram for bil type:



”Bar Chart” har til forskjell fra histogram ikke med de punktene på aksen som ikke har noen kasus. Dette kan gi litt gal forestilling av virkeligheten. Bruk derfor histogram ved kontinuerlige data. Prøv et histogram med normalkurve for en kontinuerlig variabel! Nedenfor er vist et eksempel for variabelen ”Employ” som viser antall år hos nåværende arbeidsgiver:



3.2 Deskriptiv statistikk

Vi bruker også her filen: **"Demo.sav"** og gjør følgende valg:

Analyze
Descriptive statistics
Descriptives

Her får man standard en enkel linje med statistikk pr. variabel man har definert. Du bør kun bruke variable på kontinuerlig skala. Under ser du et eksempel:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Years at current address	6400	0	56	11,56	9,938
Years with current employer	6400	0	57	10,57	9,724
Number of people in household	6400	1	9	2,35	1,468
Valid N (listwise)	6400				

Velger vi "Options" under "Descriptives" får vi nye valg. Velg nå kun en av variablene over og kryss av for ønskede parametre. Vi ser av utskriften under at vi får de samme statistiske målene som vi fikk i kapittelet om "Frequencies". (NB! "SE mean" er det samme som vi under Frequencies kalte standardfeilen til gjennomsnittet "SEM"= "Standard Error of the MEAN").

Descriptive Statistics

	N	Minimum	Maximu	Mean		Std.	Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error
Years at current address	6400	0	56	11,56	,12	9,938	,675	,061
Valid N (listwise)	6400							

3.3 Krysstabell

Vi fortsetter å bruke filen: **"Demo.sav"** og gjør følgende valg:

Analyze
Descriptive statistics
Crosstabs

Vi ønsker nå å krysstabulere eier video (kolonne) mot inntektskategori (rad). Vi velger ofte den variabelen med færrest observasjoner som kolonnevariabel.

Ut fra standard oppsett får vi følgende tabell:

Income category in thousands * Owns VCR Crosstabulation

Count

		Owns VCR		
		No	Yes	Total
Income category in thousands	Under $25	177	997	1174
	$25 - $49	75	2313	2388
	$50 - $74	2	1118	1120
	$75+	1	1717	1718
Total		255	6145	6400

Vi går nå tilbake til valgene i "Crosstabs". Dette kan vi gjøre ved å trykke på ikonet "Dialog Recall" som du finner i verktøylinjen øverst. Vi velger siste arbeidsoperasjon; "Crosstabs", og får fram bestillingene vi nettopp gjorde. Vi velger nå "Cells" og får opp en ny dialogboks. Her krysser vi av for innholdet i hver celle; for eksempel ønsker vi å se forventet verdi sammen med faktisk observert og vi krysser av "Expected" under "Counts". Under "Percentages" krysser vi av "Rows" og "Total" for å få fram hvordan inntektskategoriene er fordelt mellom de som eier og ikke eier videoutstyr. Vi trykker "Continue" og "OK".

Gjenspeiler fordelingen andelen som har og ikke har videoutstyr?

Income category in thousands * Owns VCR Crosstabulation

			Owns VCR		
			No	Yes	Total
Income category in thousands	Under $25	Count	177	997	1174
		Expected Count	46,8	1127,2	1174,0
		% within Income category in thousands	15,1%	84,9%	100,0%
		% of Total	2,8%	15,6%	18,3%
	$25 - $49	Count	75	2313	2388
		Expected Count	95,1	2292,9	2388,0
		% within Income category in thousands	3,1%	96,9%	100,0%
		% of Total	1,2%	36,1%	37,3%
	$50 - $74	Count	2	1118	1120
		Expected Count	44,6	1075,4	1120,0
		% within Income category in thousands	,2%	99,8%	100,0%
		% of Total	,0%	17,5%	17,5%
	$75+	Count	1	1717	1718
		Expected Count	68,5	1649,5	1718,0
		% within Income category in thousands	,1%	99,9%	100,0%
		% of Total	,0%	26,8%	26,8%
Total		Count	255	6145	6400
		Expected Count	255,0	6145,0	6400,0
		% within Income category in thousands	4,0%	96,0%	100,0%
		% of Total	4,0%	96,0%	100,0%

Vi ser at andelen som ikke eier videoutstyr er 4% og andelen som eier videoutstyr er 96%.

Forventet antall (”Expected Count”) regnes da ut som antall i hver inntektsgruppe multiplisert med henholdsvis 4% og 96%.. For eksempel er forventet antall som ikke har video i inntektsgruppen $25-49 lik 4%*2388=95,1. Faktisk antall (”Count”) var i dette tilfellet 75.

Tilbake i ”Crosstabs” har vi enda et hovedvalg som heter ”Statistics”. Velger vi dette kommer vi inn i ny dialogboks der vi i denne omgang krysser av for ”Chi-square”.

Dette er den mest brukte testobservatoren. Den sier noe om det er en sammenheng mellom kolonnevariabelen og rekkevariabelen. Den stiller få krav til det underliggende materialet. Vi kan for eksempel bruke nominelle variable slik som her.

Ho: Det er ingen sammenheng mellom kolonne- og rekkevariabelen; det er ingen sammenheng mellom hvorvidt man eier video og

inntektskategori. Den eventuelle forskjell i fordeling i celler mellom de ulike kolonner og rekker skyldes tilfeldigheter.

H1: Det er en sammenheng mellom kolonnevariabel og rekkevariabel; det er en sammenheng mellom hvorvidt man eier video og inntektskategori. Det finnes en reell forskjell på fordelingene mellom de ulike rekker og kolonner.

Income category in thousands * Owns VCR Crosstabulation

			Owns VCR		
			No	Yes	Total
Income category in thousands	Under $25	Count	177	997	1174
		% within Owns VCR	69,4%	16,2%	18,3%
	$25 - $49	Count	75	2313	2388
		% within Owns VCR	29,4%	37,6%	37,3%
	$50 - $74	Count	2	1118	1120
		% within Owns VCR	,8%	18,2%	17,5%
	$75+	Count	1	1717	1718
		% within Owns VCR	,4%	27,9%	26,8%
Total		Count	255	6145	6400
		% within Owns VCR	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	493,651[a]	3	,000
Likelihood Ratio	434,754	3	,000
Linear-by-Linear Association	335,462	1	,000
N of Valid Cases	6400		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 44,63.

Vi ser at vi har fått en Pearson Chi-square verdi = 493,651 og en signifikansverdi = 0,000[5]. Dette betyr at maks p-verdi er 0,0005 ved vanlige avrundingsregler. Høy Chi-square verdi angir sikre forskjeller i fordelingen av inntektsgrupper på de to kategoriene har/har ikke video.

Vi får alltid et varsel nederst i Chi-square test boksen. Det oppgir antall celler med forventet antall < 5. Dette er et eksempel på en av to tommelfingerregler når vi bruker Chi-Square test i SPSS:

1. Minimum forventet antall pr. celle i en krysstabell er 5

[5] Vi kan selvsagt redigere antall decimaler når vi er inne i output området i SPSS.

2. For tabeller som er større enn 2*2, er minimum forventet antall pr. celle =1, så lenge ikke flere enn 20% av cellene har forventet verdi < 5.

Dersom dette hadde vært et problem, kunne vi slått sammen noen av inntektsgruppene for å få resultat som er mer robuste.

3.4 Redigering og Utskrift av tabeller

Vi fortsetter å bruke filen: **"Demo.sav"** og gjør følgende valg:

Window
    Output 1 (f.eks)

Her ser vi alle utskriftene vi har produsert så langt. Til høyre ser vi en disposisjon over utskriftene. Ønsker vi å fjerne alle velger vi fra menyen:

Edit
    Select All

Så trykker du "Delete".

Vil du ta vare på utskriftene velger du fra menyen:
File
    Save As

Du får et forslag til navn med endelsen ". Spo". Dette angir at det er en "Output- fil".

Beveg deg nå til en tabell du ønsker å redigere; f.eks den første frekvenstabellen vi lagde over . Klikk en gang med musetasten og du får en enkel ramme rundt. Dobbeltklikk og du får en skravert ramme rundt. Menyen øverst har nå endret seg noe og du får opp en "Formatting toolbar 1".

Her ligger det mange muligheter, og jeg nevner kun noe få:

Marker en kolonne med prosenttall - klikk høyre musetast - velg "Cell Properties" - velg prosentformat uten desimaler.

Dobbeltklikk direkte på en tekst. Gå inn og gjør endringer i teksten, for eksempel endre til norske overskrifter.

Velg fra menyen "Format" og "Table looks". Her velger du "Academic". Tabellen får et annet utseende.

Plasser deg på en overskrift du ønsker å forklare med en fotnote. Velg fra menyen "Insert" og "Footnote".

Her finnes mange muligheter. Redigeringseksempelet over ga følgende resultat. For å få det over i dette worddokumentet klikket jeg en gang på tabellen (enkel ramme). Deretter valgte jeg fra menyen:

Edit
    Copy object (evt. Copy)

Jeg åpnet så programmet ”Word” og limte det inn på vanlig måte (CTRL+V).

Inntekts kategori

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under $25[a]	1174	18%	18%	18%
	$25 - $49	2388	37%	37%	56%
	$50 - $74	1120	18%	18%	73%
	$75+	1718	27%	27%	100%
	Total	6400	100%	100%	

a. inkluderer også sosial stønad.

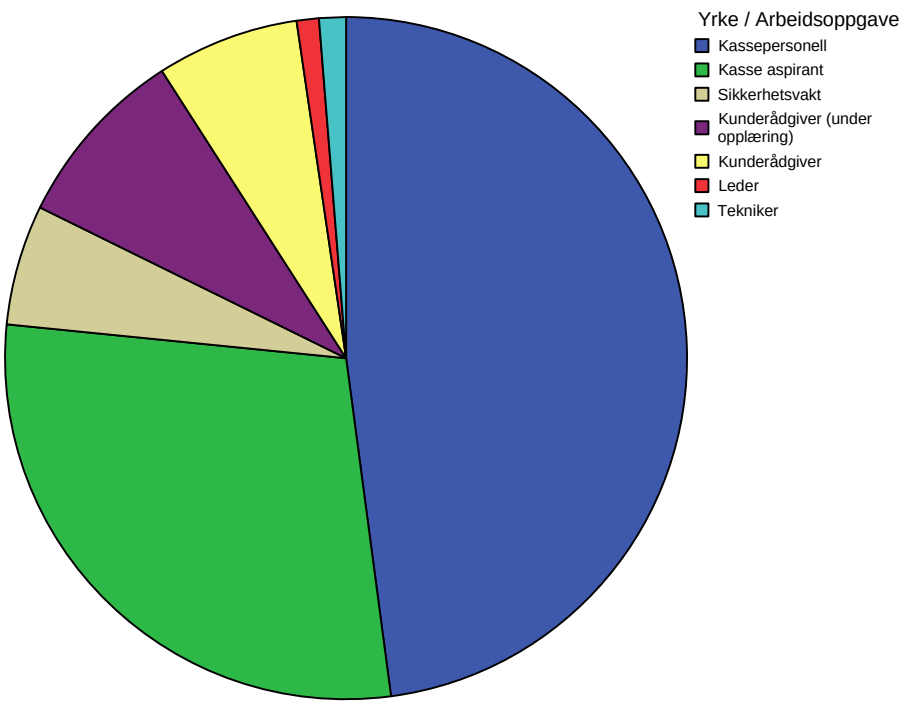
3.5 Grafikk

Vi skal nå bruke filen: **”Banknor.sav”**.

Velg fra menyen:

Graphs
    Pie
        Summaries for groups of cases

Vi lar standard valget ”N of Cases” stå. Pil inn variabelen ”Yrke” i ”Define Slices By”. Trykk ”OK”:



Skal vi endre utseende på grafen vår, så dobbeltklikker vi med venstre musetast på grafen. Vi får fram en ny boks kalt ”Chart Editor”.

Her er det mange muligheter. Jeg nevner kun et par:

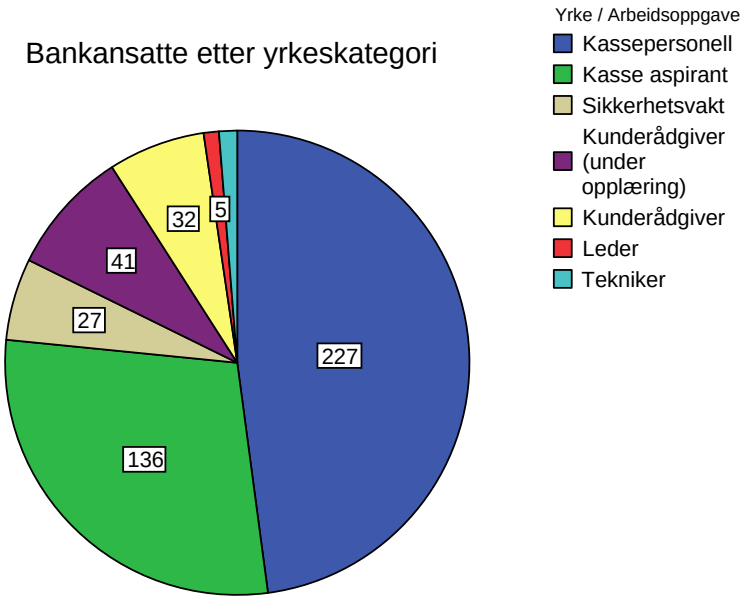
Et enkeltklikk på forklarende tekster til paistykkene gir rammer rundt hver tekst. Høyreklikk og du får fram en boks med flere valg. Her velger du ”Properties Window”. I dette vinduet kan du finstyre utseende på teksten; for eksempel velge skriftstørrelse 12.

Alle delene i utskriften er objekt. Du merker et objekt ved å venstreklikke i nærheten av eller på objektet . Du skal da få fram en ramme med svarte klosser. Merk hele paien som et objekt. Du kan nå skalere ned størrelsen på paien. Høyreklikk deretter på paien. Nå får du fram en rekke valg. Du velger ”Properties Window”. Nå kan du legge inn ”Data Labels”; dvs. antall observasjoner i hvert paistykke. Andre muligheter er at du kan slå sammen kategorier med mindre enn 5% eller fjerne kategorier.

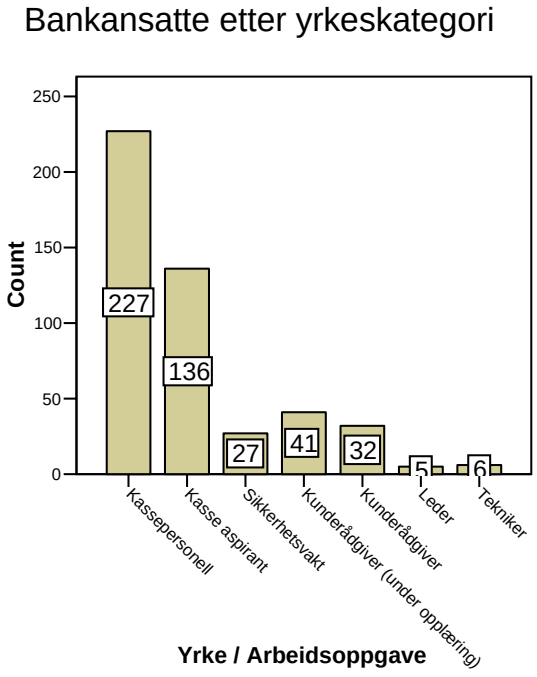
Ønsker du en tittel eller fotnote til diagrammet kan du velge fra menyen:

Chart
Add Chart Element
Text box

Jeg har gjort noen av disse endringene, og resultatet ble slik:



Ønsker du heller å bruke et stolpediagram, høyreklikker du igjen på paien. Du velger så "Change to Simple Bar" i den boksen som dukker opp:



4 Statistiske tester: gjennomsnitt

4.1 Tabeller med gjennomsnitt

Vi ønsker å se i hvilken grad gjennomsnittet til en kontinuerlig variabel varierer med ulike kategorier av respondenter.

Vi åpner filen: **”Demo.sav”**.

Vi ønsker å se om det er en forskjell på gjennomsnittprisen på bil (”Car”) mellom gifte og ugifte (”Marital”). Fra menyen velger vi:

Analyze
Compare Means
Means

Vi legger ”Car” i ”Dependent List” og ”Marital” i ”Indenpendent List” (1. undergruppe/layer 1of 1). Dette gir følgende resultat:

Report

Price of primary vehicle

Marital status	Mean	N	Std. Deviation
Unmarried	30,1812	3224	22,05847
Married	30,0748	3176	21,79590
Total	30,1284	6400	21,92692

Vi ser at gjennomsnittprisen for biler er marginalt høyere for gruppen med ugifte i denne stikkprøven.

Vi ønsker å legge inn en gruppering til; inntektskategori (”Inccat”). Vi går tilbake i dialogboksen ved å trykke på ikonet ”Dialog Recall” og velge ”Means”. I underboksen som er kalt ”Layer 1 of 1” trykker du nå på ”Next”, og velger variabelen ”Inccat” under ”Independent List”. Vi får følgende utskrift:

Report

Price of primary vehicle

Marital status	Income category	Mean	N	Std. Deviation
Unmarried	Under $25	9,3673	578	2,01810
	$25 - $49	17,6178	1228	3,60754
	$50 - $74	30,3415	552	3,56582
	$75+	61,7859	866	16,43554
	Total	30,1812	3224	22,05847
Married	Under $25	9,3876	596	2,01611
	$25 - $49	17,7693	1160	3,55358
	$50 - $74	30,1845	568	3,64916
	$75+	61,2269	852	16,30229
	Total	30,0748	3176	21,79590
Total	Under $25	9,3776	1174	2,01626
	$25 - $49	17,6914	2388	3,58148
	$50 - $74	30,2619	1120	3,60757
	$75+	61,5087	1718	16,36721
	Total	30,1284	6400	21,92692

Den siste utskriften tyder på at det er en betydelig forskjell i bilprisen mellom inntektsgruppene. Men er forskjellen et resultat av utvalgsusikkerhet eller er det statistisk signifikante forskjeller?

4.2 Tester av to gjennomsnitt- uavhengige utvalg

Vi fortsetter å bruke filen: **"Demo.sav"**.

Våre data viser en noe høyere gjennomsnittspris for biler i gruppen ugifte. Vi betrakter nå disse dataene som en stikkprøve (utvalg) av en større populasjon; der populasjonen for eksempel er innbyggere i USA. Er forskjellen i bilpris mellom gifte og ugifte i denne stikkprøven så markert at vi kan påstå at den ikke skyldes tilfeldigheter?

Det vil si vi ønsker å teste følgende nullhypotese Ho mot den alternative hyposten H1:

Ho: Ugifte og gifte i USA kjøper like dyre biler i snitt
H1: Ugifte og gifte i USA kjøper ikke like dyre biler i snitt

$$H_0 : \mu_{ugift} = \mu_{gift} = \mu$$
$$H_1 : \mu_{ugift} \neq \mu_{gift} \neq \mu$$

For å teste dette bruker vi den uavhengige[6] t-testen. Formelen under gir 95% konfidensintervall for differansen. Hvis dette intervallet innholder 0, beholder vi nullhypotesen på 5% signifikansnivå.

[6] Utvalgene vi sammenligner er gifte og utgifte i fila demo.sav. Disse utvalgene er uavhengige. En avhengig t-test vil typisk være å gjøre parvise sammenligninger på samme

$$(\mu_1 - \mu_2) = (\overline{X}_1 - \overline{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Bruk av t-testen krever at vi har kontinuerlige data (prisen på bil nummer 1 i husholdningen:"Car") og at vi kan forutsette at gjennomsnittlig bilpris for de to gruppene gift og ugift er rimelig normalfordelt. Vi har grunn til å tro at bilpris generelt ikke er normalfordelt i de to gruppene, men at de vil ha en skjevhet da bilprisen er begrenset nedad til null mens noen kjøper ekstremt dyre biler. Utskrift av histogram for variabelen "Car" samt estimerte parametere for skjevhet, kurtose og median underbygger dette. Likevel kan vi forutsette tilnærmet normalitet for gjennomsnittsprisen på bil til de to gruppene - hvorfor?[7]

Videre må vi forutsette at stikkprøven (spørreundersøkelsen) er representativ for gifte og ugifte i USA. Dette er tilfelle viss hver og en av respondentene er tilfeldig valgt ut blant alle innbyggerne i USA. For undervisningens skyld, forutsetter vi derfor dette nå.

Vi må og forutsette at de ulike kasus er statistisk uavhengige. For eksempel at det ikke er gjennomført gjentatte undersøkelser på en og samme person.

Til slutt må vi forutsette at de to gruppene; bilpris gifte og bilpris ugifte, har samme underliggende varians i populasjonene. Tabellen nedenfor har en egen test for dette. Hvis Levenes test er signifikant ($p<0,05$), så er variansene i utrykkene for menn og kvinners lønn forskjellige og vi bruker nedre linjes tall.

Vi velger fra menyen:

Analyze
Compare Means
Independent Samples t-test

I dialogboksen legger vi inn "Car" som testvariabel og "Marital" som "Grouping variabel". Vi klikker på "Define groups" og skriver inn 0 (= ugift) for gruppe 1 og 1 (= gift) for gruppe 2. Viss du er i tvil om verdiene, høyreklikk på variabelen "Marital" i dialogboksen. Vi får da følgende utskrift:

---

sett personer. For eksempel sammenligne salget av en vare i utvalgte butikker før og etter en markedskampanje. Differansen blir så analysert for å se om den er signifikant forskjellig fra null.

[7] Sentralgrenseteoremet: Selv om populasjonen ikke er normalfordelt vil gjennomsnittsverdiene i stikkprøver bli tilnærmet normalfordelt. Ved svært skjeve fordelinger i populasjonen, kreves et høyt antall observasjoner.

T-Test

Group Statistics

	Marital status	N	Mean	Std. Deviation	Std. Error Mean
Price of primary vehicle	Unmarried	3224	30,1812	22,05847	,38849
	Married	3176	30,0748	21,79590	,38675

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Price of primary vehicle	Equal variances assumed	,512	,474	,194	6398	,846	,1064	,54823	-,96832	1,18111
	Equal variances not assumed			,194	6397,941	,846	,1064	,54818	-,96822	1,18101

Vi ser at ugifte i snitt betaler 1064 dollar mer for sin bil (mean difference: 0,1064).

På bakgrunn av tabellen over beholder vi nullhypotesen om at gjennomsnittprisen på bil for gifte og ugifte er lik. Vi kan gjøre det på en av tre måter:

- ⇨ Siden ***konfidensintervallet*** innholder null, beholder vi nullhypotesen om lik bilpris i snitt.
- ⇨ Dette kan vi alternativt se direkte ved å studere **Sig. (2-tailed)** som med 3 desimaler er 0,846. Gitt at nullhypotesen er rett, så er det 84,6% sannsynlig at du i et tilfeldig utvalg får en forskjell på 0,1064 eller mer (i absoluttverdi).
- ⇨ Den ***observerte t*** = ”Mean Difference”/”Std. Error Difference” = 0,194 ligger godt innenfor den kritiske t-verdien vi bruker ved 5% signifikansnivå; 1,96.

Et par kommentarer:

Den kritiske t-verdien i et tosidig 95% konfidensintervall er 1,96 ved så store utvalg som vi har her (6400 observasjoner). Sjekk tabell i læreboka.

Fra utskriften over ser vi at SE (standardfeilen til gjennomsnittlig differanse) er 0,54823. Vi kan da selv regne ut konfidensintervallet som: 0,1064 +/- 1,96*0,54823 => laveste verdi : -0,9681.. og høyeste verdi 1,1809. Avviket fra konfidensintervallet gitt over skyldes kun avrunding.

En p-verdi eller Sig (2-tailed) på mindre enn 5% ville ført til at vi forkastet nullhypotesen.

4.3 Tester av to gjennomsnitt - avhengige utvalg

Vi skal nå studere en situasjon der vi har to målinger av samme variabel pr kasus. Hent fram fila: **"Banknor.sav"**. Vi skal her studere forskjell i startlønn og sluttlønn for hver enkelt ansatt og foreta en såkalt parret t-test for å finne om forskjellen i startlønn og sluttlønn er signifikant[8].

Det vil si vi ønsker å teste følgende nullhypotese Ho mot den alternative hyposten H1:

Ho: Bankansatte i USA tjener det samme når de begynner som når de slutter i sin stilling
H1: Bankansatte i USA har en annen sluttlønn enn begynnerlønn

$$H_0 : \mu_{start} = \mu_{slutt} = \mu$$
$$H_1 : \mu_{start} \neq \mu_{slutt} \neq \mu$$

For å teste dette bruker vi den avhengige t-testen. Formelen under gir 95% konfidensintervall for differansen. Hvis dette intervallet innholder 0, beholder vi nullhypotesen på 5% signifikansnivå.

$$\Delta = \overline{D} \pm t_{0,025} \frac{s_D}{\sqrt{n}}$$

Vi foretar følgende valg fra menyen:

Analyze
Compare Means
Paired Samples T-test

I dialogboksen velger vi variabelen "Slutlønn". Når vi merker denne, går den ned i "Current Selection" boksen som variabel 1. Vi merker tilsvarende variabelen "Startløn", og denne går inn som variabel 2. Bruker vi piltasten til å flytte over variablene, får vi startlønn - sluttlønn som variabel over i "Paired Variabel" boksen. Vi trykker OK og får følgende output:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Start-lønn	6806,43	474	3148,26	144,60
	slut-lønn	13767,83	474	6830,26	313,72

[8] Eksempelet er dårlig. Utvalget er hentet fra en og samme bedrift, og sluttlønn (lønn på måletidspunktet) bør jo være større enn på det tidspunktet man ble ansatt dersom man ikke har skiftet arbeidstype innenfor samme bedrift. Gode eksempler ville vært effekten av en medisin (måling blodtrykk samme personer før og etter medisinering) eller effekten av markedsføringskampanje (måling samme salg utvalgte varer før og etter kampanjen).

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Start-lønn & slut-lønn	474	,880	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Start-lønn - slut-lønn	-6961,39	4325,85	198,69	-7351,82	-6570,96	-35,036	473	,000

Vi har her fått 474 observasjoner av den nye variabelen startlønn - sluttlønn. De 474 tallparene som ligger bak denne variabelen, har en korrelasjon på 88%. Gjennomsnittlig startlønn for alle 474 er 6.806,43. Gjennomsnittlig sluttlønn er 13.768,83.

På et 5% signifikansnivå kan vi forkaste nullhypotesen om ingen forskjell startlønn og sluttlønn fordi:

- ⇨ Konfidensintervallet ikke innholder null
- ⇨ Absoluttverdien av observert t er langt større en kritisk t-verdi (1,96-1,98)
- ⇨ P-verdien til en tosidig test (Sig. (2-tailed)) er mindre enn 0,05.

Vi kan regne ut konfidensintervallet manuelt ut fra formelen over:

-6.961,39 +/- 1,97*(4325,85/rot(474))
-6.961,39 +/- 391,42
nedre grense: -6.570
øvre grense: -7.353

Avvik fra tabellen over skyldes avrunding

4.4 Tester av flere gjennomsnitt (ANOVA)

Enveis ANOVA gir en analyse av variansen til en avhengig variabel forklart ved en faktor eller uavhengig variabel. Den avhengige variabelen må ha et kontinuerlig målenivå, mens faktoren må ha et ordinalt målenivå. Variansanalysen ANOVA brukes til å teste hypotesen om at flere gjennomsnitt er like:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = ... = \mu$$
$$H_1 \neq \mu_1 \neq \mu_2 \neq \mu_3 \neq ... \neq \mu$$

Hent fram fila: **"Demo.sav"**.Vi ønsker nå å teste hypotesen om at gjennomsnittpris for bil for ulike inntektskategorier i USA er lik. Vi

forutsetter også her at utvalget i datafilen demo.sav er representativ for alle innbyggere i USA. Vi vet at det er forskjeller i gjennomsnittlig bilpris mellom inntektsgruppene i dette utvalget, men er forskjellene så betydelige at de ikke kan forklares ut fra tilfeldigheter knyttet til dette utvalget?

Vi velger fra menyen:

Analyze
Compare Means
One-Way Anova

I dialogboksen velger vi "Car" (kontinuerlig variabel) som dependent variabel og "inccat" som factor. Antall inntektskategorier bestemmer da antall gjennomsnitt vi sammenligner. Vi får følgende utskrift:

ANOVA

Price of primary vehicle

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2566663	3	855554,483	10731,559	,000
Within Groups	509909,7	6396	79,723		
Total	3076573	6399			

Tabellen over kan vi lese slik:

"Sum of Squares" uttrykker variasjonen (ikke variansen) i tallmaterialet. Av en total variasjon på 3.076.573 er 2.566.663 forklart gjennom variasjon mellom gjennomsnittsprisen på bil til de ulike inntektsgruppene, mens 509.909 er et resultat av tilfeldig variasjon (= uforklart).

"Df" er antall frihetsgrader. Antall frihetsgrader for den forklarte variansen er lik antall grupper i faktoren – 1. Her har vi 4 yrkesgrupper i faktoren "Inccat" og får 3 frihetsgrader. Antall frihetsgrader for den uforklarte variansen er lik antall observasjoner – antall grupper i faktoren. Vi får 6400-4=6396.

"Mean Square" uttrykker variansen i stikkprøven. Øverst har vi variansen mellom gjennomsnittsprisen på bil til hver inntektsgruppe og total gjennomsnittspris multiplisert med et veid snitt av antall observasjoner i hver inntektsgruppe. Nederst har vi variansen mellom bilpriser og gjennomsnitts bilpris for hver inntektsgruppe. Er nullhypotesen sann, skal disse to utrykkene være like i populasjonen. Dette betyr at vi forventer en F i nærheten av 1 i vår stikkprøve. Kritisk verdi F avhenger av antall frihetsgrader og vårt krav til signifikansnivå. F = 855.554/79,7= 10734. (avvik skyldes avrunding). Fra tabell finner vi at kritisk verdi F på 5% signifikansnivå er 2,10 (df går mot uendelig). Vi forkaster dermed nullhypotesen om ingen forskjell i lønn mellom inntektsgruppene.

Vi får det samme resultatet ved å betrakte p-verdien (= Sig.). Gitt at nullhypotesen er sann, er det mindre enn 0,0005[9] sannsynlighet for å få de observerte tallene i en stikkprøve. Vi forkaster dermed nullhypotesen.

Analysen over kan utvides ved å krysse av på noen nye alternativ i dialogboksen. Vi går tilbake fra outputfilen ved å trykke på ikonet "Dialog Recall". Vi får opp "One way ANOVA" dialogboksen. Vi går så inn i dialogboksen "Options" og klikker på de tre valgene "Means plot", "Descriptive" og "Homogenity of variance test". Vi får nå opp følgende utskrifter i tillegg til ANOVA tabellen som vi har vist over:

Descriptives

Price of primary vehicle

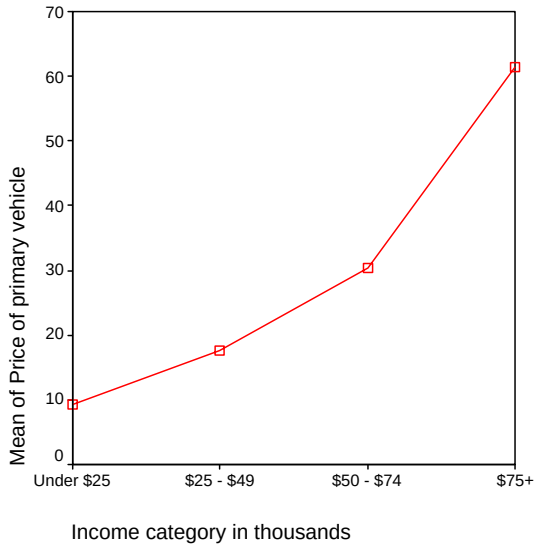
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Under $25	1174	9,3776	2,01626	,05885	9,2621	9,4931	4,20	13,00
$25 - $49	2388	17,6914	3,58148	,07329	17,5477	17,8351	11,90	25,70
$50 - $74	1120	30,2619	3,60757	,10780	30,0504	30,4734	23,30	38,20
$75+	1718	61,5087	16,36721	,39488	60,7342	62,2832	36,20	99,90
Total	6400	30,1284	21,92692	,27409	29,5911	30,6657	4,20	99,90

Test of Homogeneity of Variances

Price of primary vehicle

Levene Statistic	df1	df2	Sig.
2970,792	3	6396	,000

Means Plots



[9] Du kan få fram den eksakte p-verdien ved å endre tall desimaler på utskriften når du er inne i outputfilen til SPSS.

Den øverste tabellen viser bl.a. konfidensintervallet for gjennomsnittlig bilpris for hver inntektsgruppe. Vi har alt forkastet nullhypotesen om at gjennomsnittlig bilpris i de ulike inntektskategoriene er lik i USA. Dette blir her demonstrert ved at ingen av konfidensintervallene er overlappende. Den nedre grensen for gjennomsnittlig bilpris for en i inntektsgruppe 4; 60,73, ligger over den øvre grensen til gjennomsnittlig bilpris i inntektsgruppe 3; 30,47 .

Under har vi en test for om variansen til bilprisen for de 4 yrkesgruppene er lik. Dette er en forutsetning for å kjøre enveis ANOVA. Vi ser at nullhypotesen om lik varians blir forkastet ved 5% signifikansnivå.

Til slutt har vi en oversikt over gjennomsnittlig bilpris for de 4 inntektsgruppene i vår stikkprøve.

Vi har nå fått slått fast at det eksisterer en differanse mellom gjennomsnittlig bilpris for de 4 inntektsgruppene i USA. Men hvilke gjennomsnitt er det som er signifikant forskjellige? Her finnes det et uttall med tester. Vi går tilbake til dialogboksen "One way ANOVA "og klikker på "Post Hoc". I dialogboksen som åpner seg kan vi for eksempel velge "LSD", "Scheffe" eller "Tamphanes T2". De to første forutsetter lik varians. Den siste kan brukes når forutsetningen om lik varians for de ulike yrkesgruppene ikke holder - slik som i dette eksempelet. Vi får følgende utskrift:

Multiple Comparisons

Dependent Variable: Price of primary vehicle

Tamhane

(I) Income category in thousands	(J) Income category in thousands	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Under $25	$25 - $49	-8,3138*	,09399	,000	-8,5613	-8,0664
	$50 - $74	-20,8843*	,12281	,000	-21,2078	-20,5608
	$75+	-52,1311*	,39924	,000	-53,1826	-51,0795
$25 - $49	Under $25	8,3138*	,09399	,000	8,0664	8,5613
	$50 - $74	-12,5705*	,13035	,000	-12,9137	-12,2272
	$75+	-43,8173*	,40162	,000	-44,8751	-42,7594
$50 - $74	Under $25	20,8843*	,12281	,000	20,5608	21,2078
	$25 - $49	12,5705*	,13035	,000	12,2272	12,9137
	$75+	-31,2468*	,40933	,000	-32,3248	-30,1688
$75+	Under $25	52,1311*	,39924	,000	51,0795	53,1826
	$25 - $49	43,8173*	,40162	,000	42,7594	44,8751
	$50 - $74	31,2468*	,40933	,000	30,1688	32,3248

*. The mean difference is significant at the .05 level.

Tabellen viser at når vi sammenligner gjennomsnittlig bilpris for to inntektsgrupper om gangen, så er gjennomsnittsprisen for de alle de 4 ulike gruppene signifikant forskjellig.

5 Korrelasjon

I visse situasjoner kan det være aktuelt å se på graden av samvariasjon mellom ulike variable. Dette måler vi med korrelasjonskoeffisienten R (for populasjonen) eller r (for utvalg).

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}$$

eller

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_X s_Y}$$

Pearsons korrelasjonskoeffisient er definert slik at

$$-1 \le r \le 1$$

Hvis høy verdi av X går sammen med høy verdi av Y og omvendt, vil r ligge nær +1. Hvis høy verdi av X følges av lav verdi av Y og omvendt, vil r ligge nær -1. Hvis X og Y beveger seg uavhengig av hverandre er r 0.

For å bruke Pearsons korrelasjonskoeffisient må dataene følge en kontinuerlig skala. Skal vi foreta statistisk testing på korrelasjonskoeffisienten, skal bakenforliggende data være tilnærmet normalfordelt. Spearmans rangordningskoeffisient kan brukes på variable som er på ordinal skala. Korrelasjonskoeffisient skal aldri beregnes på variable på nominal skala.

Korrelasjonskoeffisienter bør ikke regnes ut uten at vi samtidig tar ut et enkelt skatterplott på de aktuelle variable, fordi underliggende sammenhenger kan bli borte ved at en ved bruk av korrelasjonskoeffisienter kun fanger opp vanlige lineære sammenhenger.

Vi henter fram filen: **”Banknor.sav”**. Vi velger fra menyen:

Analyze
    Correlate
        Bivariate

I dialogboksen velger vi de kontinuerlige variablene ”Slutlønn”, ”Startløn”, ”Alder” og ”Utdantid”. Vi klikker OK og får:

Correlations

		Start-lønn	slut-lønn	Alder	Utdannelsestid i år
Start-lønn	Pearson Correlation	1,000	,880**	-,011	,633**
	Sig. (2-tailed)	,	,000	,811	,000
	N	474	474	474	474
slut-lønn	Pearson Correlation	,880**	1,000	-,146**	,661**
	Sig. (2-tailed)	,000	,	,001	,000
	N	474	474	474	474
Alder	Pearson Correlation	-,011	-,146**	1,000	-,281**
	Sig. (2-tailed)	,811	,001	,	,000
	N	474	474	474	474
Utdannelsestid i år	Pearson Correlation	,633**	,661**	-,281**	1,000
	Sig. (2-tailed)	,000	,000	,000	,
	N	474	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

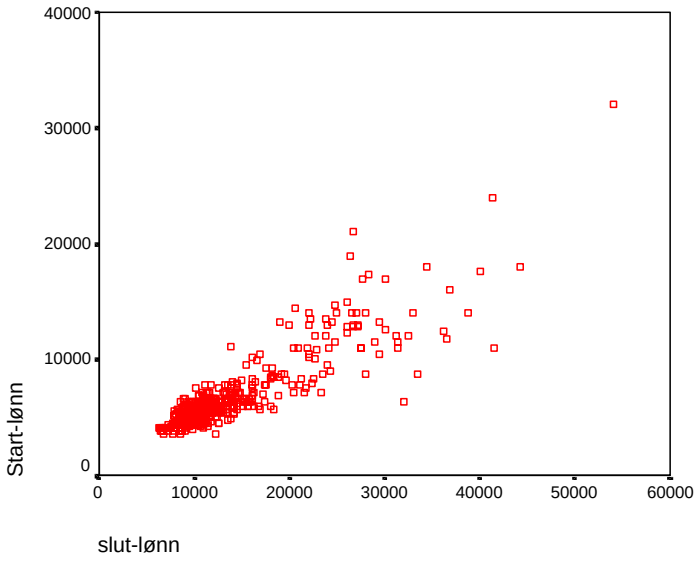
Koeffisientene på diagonalen blir alle lik 1 da alle variable er fullstendig korrelert med seg selv. Tallene nede til venstre under diagonalen er et speilbilde av tallene over diagonalen. Vi ser at den sterkeste korrelasjonen er mellom sluttlønn og startlønn. Det er også en sterk korrelasjon mellom sluttlønn og utdannelsestid. For alle disse sammenhengene er korrelasjonen signifikant på 5% nivået. Det er liten korrelasjon mellom alder og sluttlønn.

Vi lager et plott mellom variablene ”Slutlønn” og ”Startløn” ved å velge følgende fra menyen:

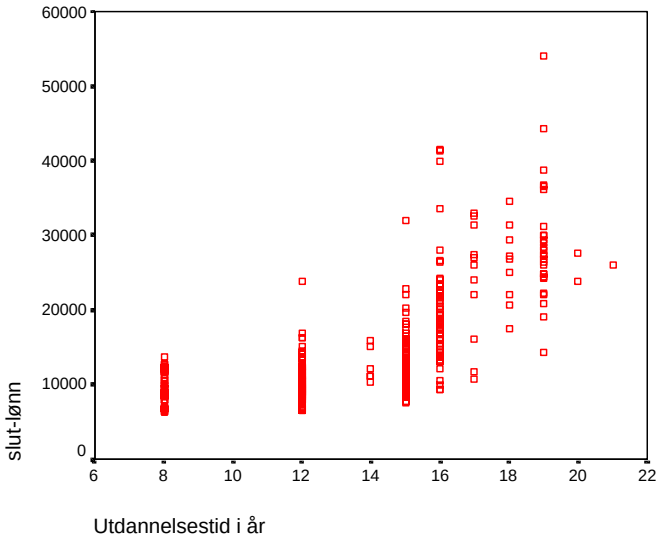
Graph
Scatter

I dialogboksen velger vi Simple og trykker Define. Vi velger ”Startløn” og ”Slutlønn” og trykker OK. Diagrammet som følger viser en klar sammenheng mellom startlønn og sluttlønn. Dette bekreftes av korrelasjonsmatrisen over der r =0,88.

En korrelasjonsmatrise sier ingen ting om hva som er årsak og hva som er virkning. Ingen variabel er spesifisert som avhengig og ingen som uavhengig.



Vi lager et tilsvarende scatter plot for variablene ”Slutlønn” og ”Utdantid”:



Vi ser en klar sammenheng mellom utdannelsestid og sluttlønn, men er den lineær?

Vi vil videre se på ***partielle korrelasjoner***. Dette er korrelasjoner mellom bestemte variable i det vi korrigerer for ulike nivå av andre variable. Vi ser på den partielle korrelasjonen mellom utdannelsestid og sluttlønn korrigert for startlønn. Vi velger fra menyen:

Analyze
Correlate
Partial

I dialogboksen legger vi inn ”Utdantid” og ”Slutlonn” i boksen for ”Variables” og ”Startlon” i boksen ”Controlling for”. Vi får resultatet[10]:

- - - P A R T I A L C O R R E L A T I O N C O E F F I C I E N T S - - -

Controlling for.. STARTLØN

	UTDANTID	SLUTLØNN
UTDANTID	1,0000	,2810
	(0)	(471)
	P= ,	P= ,000
SLUTLØNN	,2810	1,0000
	(471)	(0)
	P= ,000	P= ,

(Coefficient / (D.F.) / 2-tailed Significance)

" , " is printed if a coefficient cannot be computed

Vi ser at korrelasjonen mellom sluttlønn og utdannelsestid er sunket fra 0,66 til 0,28 idet vi justerer for ulike nivå av startlønn. Prøv å finne en fornuftig forklaring på dette!

En justering for en tredje faktor kan gjøre at korrelasjonene mellom to faktorer går både opp og ned.

[10] Utskriften som følger er fra SPSS versjon 11.5.

6 Regresjon

6.1 Regresjonsanalyse; arbeidsmåte

Vi skal nå arbeide med spesifisering av regresjonsmodeller og tolking av analyseresultatene. Innledningsvis bør dere merke dere:

1. Gjør dere alltid opp en á priori mening om hvilke variable som forklarer den uavhengige variabelen Y.

2. Formuler nullhypoteser og alternative hyposteser for hver sammeheng; eller sagt med andre ord for hver variabel og dens beta.

Dette skal dere gjøre før dere har gjort et eneste tastetrykk på SPSS. Hypotesene skal være basert på formulert teori. Når dere så har gjort analysen på SPSS bør dere merke dere følgende før dere konkluderer:

1. En høy signifikanssannsynlighet[11] vil normalt bidra til at du beholder nullhypotesen om ingen sammenheng mellom X og Y. La imidlertid ikke SPSS alene styre om du velger å beholde en uavhengig variabel X i regresjonsmodellen. Har du en sterk á priori grunn til å inkludere en uavhengig variabel, så kan det likevel være rett å inkludere den i analysen, spesielt hvis fortegnet på koeffisienten er rett. Kanskje skyldes den høye p-verdien at du har en liten stikkprøve[12] Den høye p-verdien kan heller ikke tolkes som at vi har bevist at det ikke er sammenheng mellom X og Y, vi har bare ikke hatt gode nok bevis for det motsatte.

2. Er den høye signifikanssannsynligheten koblet med en litt usikker á priori oppfatning av sammenhengen kan vi ofte fjerne den uavhengige variabelen fra modellen. Vær imidlertid obs. på indirekte effekter!

3. Uventede resultater kan gjøre at vi stiller spørsmål ved om samtlige viktige forklarende variable er inkludert. Er modellen velspesifisert, eller innholder residualene systematisk støy som vi bør undersøke nærmere? Det kan og være at vi stiller spørsmål ved variablenes validitet. Måler variabelen det den er ment å måle? Hvis ikke, bør variabelen fjernes fra modellen.

Vanlige antakelser i regresjon:

[11] Signifikanssannsynlighet = p-verdi =sig. i SPSS tabellene. Dette er sannsynligheten for å få stikkprøvens t-verdi eller høyere (absoluttverdi) under forutsetning av at nullhypotesen om ingen sammenheng X og Y er rett. Er sannsynligheten svært liten; normalt under 5%, forkaster vi nullhypotesen. Dvs. vi mener at det er sannsynliggjort at det er en sammenheng mellom den uavhengige variabelen X og Y. Beta er forskjellig fra null.

[12] En liten stikkprøve gir et bredere konfidensintervall og større sannsynlighet for at du beholder nullhypotesen, alt annet like.

1. Regresjonsmodellen er linær:
   $Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_n X_{ni} + \varepsilon_i$

2. Forklaringsvariablene er gitte konstanter (evt. eksogent gitt)

3. Fravær av multikollinearitet

4. Uavhengighet; dvs. fravær av autokorrelasjon

5. Homoskedastisitet

6. Normalfordelte restledd

Forutsetning 1-2 sikrer forventningsretthet. Forutsetning 1- 5 sikrer både forventningsretthet og effisiens. Når disse forutsetningene er tilfredsstilt vil det å bruke minste kvadraters metode gi estimerte parametere (alfa, beta, varians etc.) som er forventningsrette og har minst varians av alle mulige metoder for estimering. Dette resultatet er svært sentralt i statistikkfaget, og kalles Gauss Markow teoremet eller BLUE (best linear unbiased estimators). Forutsetning 6 er strengt tatt ikke nødvendig for å kunne estimere gjennomsnitt, beta, standardavvik etc. Men, skal vi foreta hypotesetesting eller danne konfidensintervall, er vi nødt til å anta noe om fordelingen til det stokastiske restleddet. Når vi antar at restleddet er normalfordelt så kan vi for en stikkprøve bruke testobservatøren;

$$t = \frac{\hat{\beta} - \beta_0}{s_\beta},$$

til å teste hyposteser og danne konfidensintervall.

6.2 Oppbygging av en multippel regresjonsmodell

6.2.1 Á priori oppfatninger av kausale sammenhenger

Vi vil ta utgangspunkt i fila **"Banknor.sav"**. Filens opphav er ukjent, men jeg forutsetter at utvalget er hentet fra amerikansk bankvesen, og at de gir et representativt bilde av lønnsvilkår etc. for denne næringen i USA. Jeg antar videre at beløp som startlønn, sluttlønn etc. er oppgitt i dollar årslønn.

I det følgende skal vi forsøke å analysere variasjonen i den kontinuerlige variabelen sluttlønn. Dette er lønnen til den enkelte ansatte på tidspunktet

for spørreundersøkelsen. Før jeg ser nærmere på dataene[13] har jeg følgende á priori antakelse om hvilke faktorer som kan tenkes å påvirke sluttlønnen:

Relevant arbeidserfaring påvirker trolig lønnen positivt alt annet like. Lønnsnivået gjenspeiler både at du utfører arbeidet ditt bedre og at lønnssystemet normalt er bygget opp slik at de fanger opp effekten av ansiennitet. Et mål på relevant arbeidserfaring kan være alder.

Nullhypotese - Ho: $\beta = 0$
Ingen sammenheng mellom relevant arbeidserfaring og lønn.

Alternativ hypotese - H1: $\beta > 0$.
Økt relevant arbeidserfaring gir økt lønn.

Relevant utdanning påvirker trolig lønnen positivt alt annet like. Et mål på dette kan være antall år utdanning etter videregående.

Nullhypotese - Ho: $\beta = 0$.
Ingen sammenheng mellom utdanning og lønn..

Alternativ hypotese - H1: $\beta > 0$.
Økt relevant utdanning. gir økt lønn.

Kjønn påvirker kanskje lønnen alt annet like. Her er jeg mer usikker, og i min alternative hypotese tar utgangspunkt i en forskjell i gjennomsnittslønn for de to gruppene i stikkprøven. Variabelen kjønn i datafilen "Banknor.sav" er utformet slik at den fungerer som en dummy der mann=0 og kvinne=1.

Nullhypotese - Ho: $\beta = 0$.
Kjønn forklarer ikke lønnsforskjeller i den amerikanske banksektoren.

Alternativ hypotese - H1: $\beta < 0$
Kvinner i den amerikanske banksektoren tjener mindre enn menn.

Yrkeskategori i banksystemet påvirker trolig lønnen. Det er opplagt at en leder tjener mer enn en som arbeider i kassen alt annet like. Forskjeller mellom andre yrkeskategorier vil avhenge litt av hvor godt yrkesgruppene er definert mht til forskjeller i ansvarsnivå og kompleksitet arbeidsoppgaver. Jeg bruker eksisterende yrkeskategorier som mål på denne variabelen og benytter 6 dummyvariable for å teste forskjellene mellom de 7 yrkesgruppene.

Nullhypotese - Ho: $\beta_{D_i} = 0$ for hver dummyvariabel.

[13] I virkeligheten ville disse antakelsene/teoriene ofte ha styrt utformingen av spørreundersøkelsen.

Det er ingen forskjell i lønnsnivå mellom yrkesgruppene i den norske banksektoren.

Alternativ hypotese - H1: $\beta_{D_i} \neq 0$ for hver dummyvariabel.
Forskjell i yrkeskategori gir forskjell i lønnsnivå i den norske banksektoren.

Jeg vil nå gradvis inkludere én og én av variablene ovenfor. Dette gjør jeg for å få fram en del poeng om hvordan forklaringskraften til modellen og koeffiisentene til de ulike variablene endres ved å stadig legge til nye uavhengige variable.

6.2.2 Case 1

Vi starter med en enkel lineær regresjon mellom sluttlønn og alder:

Analyze
Regression
Linear

I dialogboksen velger vi ”Slutlon” i boksen ”Dependent”og ”Alder” i boksen ”Independent”. Den avhengige variabelen må ha en kontinuerlig skala, mens de avhengige kan ha både nominell skala (kategorier), ordinal skala (rangering) og kontinuerlig skala. Vi bruker dummy-variable for variable med en nominell eller ordinal skala.

Vi får følgende resultat:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Alder[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: slut-lønn

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,146[a]	,021	,019	6764,32

a. Predictors: (Constant), Alder

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,70E+08	1	469794808,1	10,267	,001[a]
	Residual	2,16E+10	472	45756026,40		
	Total	2,21E+10	473			

a. Predictors: (Constant), Alder

b. Dependent Variable: slut-lønn

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16911,899	1029,226		16,432	,000
	Alder	-84,550	26,386	-,146	-3,204	,001

a. Dependent Variable: slut-lønn

Vi ser at forklaringskraften til modellen er svært liten; ”R Square” = 0,021. Forskjeller i alder forklarer mindre enn 2% av variasjonen i lønn på tidspunktet for spørreundersøkelsen. ”Adjusted R Square” er litt mindre og det kommer av at vi her har justert for antall frihetsgrader.

Dette kan vi alternativt se ved å gå inn i ANOVA tabellen og dele ”SS Regression” på ”SSTotal”; 0,047*10^10/2,21*10^10=0,021. Samtidig ser vi fra ANOVA tabellen at til tross for at regresjonen har liten forklaringskraft, så er det en signifikant sammenheng mellom alder og lønn. Dette kan vi se ved å observere F verdien på 10,267 og sammenligne den med kritisk F-verdi i tabell. Eller vi ser det direkte av den lave p-verdien i ANOVA tabellen; Sig. = 0,001.

I tabellen ”Coeffisients” ser vi at regresjonskoeffisienten til alder er -84,550. Det betyr at en bankansatt som er ett år eldre enn en annen, vil tjene knapt 85 mindre i gjennomsnitt. Dette gjelder for denne stikkprøven. Skal vi generalisere bør vi supplere dette estimatet med et konfidensintervall. Konfidensintervallet kan vi regne ut på bakgrunn av informasjonen i tabellen og kritisk t-verdi[14] til; -84,550 +/- 1,96*26,386. Siden intervallet ikke innholder null er sammenhengen signifikant; dvs. vi kan forkaste nullhypotesen om at det ikke er noen sammenheng mellom alder og lønn. Dette kan vi og se ved å observere p-verdien til regresjonskoeffisienten som her er lik 0,001.

Men vent nå litt! Har virkelig alder en negativ innvirkning på lønn alt annet like?

[14] For stikkprøver > 120 observasjoner bruker vi Z fordelingen.

Fra læreboka vet vi at vi bør utvide en enkel regresjonsmodell til å inkludere flere variable som påvirker sluttlønn. Slik kan vi rendyrke effekten av alder på sluttlønn ved at koeffisienten b for alder tolkes som effekten av å være et år eldre gitt at de andre forklarende variablene holdes konstant. I tillegg ønsker vi selvsagt å øke forklaringskraften til modellen ved å inkluderer flere relevante variable.

6.2.3 Case 2

Vi velger:

Analyze
Regression
Linear

Denne gangen legger vi inn både alder og utdannelsestid som uavhengige variable i boksen "Independent". Vi velger metode "Enter", [15], og får følgende resultat:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Utdannelsestid i år, Alder[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: slut-lønn

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,662[a]	,438	,436	5131,10

a. Predictors: (Constant), Utdannelsestid i år, Alder

[15] Velger vi method = stepwise, vil SPSS forkaste variabler som ikke er signifikante. Vi ønsker å gjøre disse vurderingene selv, og ikke overlate de til statistikkprogrammet. Derfor velger vi standard innstilling der samtlige spesifiserte forklaringsvariable inngår i modellen i den regresjonsmodellen som lages.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,67E+09	2	4833043061	183,569	,000[a]
	Residual	1,24E+10	471	26328138,32		
	Total	2,21E+10	473			

a. Predictors: (Constant), Utdannelsestid i år, Alder

b. Dependent Variable: slut-lønn

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-8644,578	1574,608		-5,490	,000
	Alder	24,913	20,855	,043	1,195	,233
	Utdannelsestid i år	1592,550	85,211	,673	18,689	,000

a. Dependent Variable: slut-lønn

Ved å inkludere også utdanning som forklaringsvariabel øker forklaringskraften markert fra om lag 2% til om lag 44%. Samtidig ser vi at mens det er en signifikant sammeheng mellom utdanning og lønn, så er det nå ikke lenger en signifikant sammenheng mellom alder og lønn. Koeffisienten foran alder er også endret fra negativ til positiv; en bankansatt i denne stikkprøven som er ett år eldre enn en annen vil i gjennomsnitt tjene 25 mer når vi korrigerer for lengde på utdanningen.

Resultatet over kan imidlertid få oss til å vurdere om vi bør ha med alder som uavhengig variabel når den ikke har en signifikant innvirkning på lønn. Vi kommer tilbake til dette når vi har inkludert samtlige variable i modellen.

6.2.4 Case 3

Vi velger nå å utvide modellen til å inkludere en kategorisk variabel; kjønn. Denne blir inkludert som en dummyvariabel der mann=0 og kvinne=1. Vi får følgende resultat:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Ansattes kjønn, Alder, Utdannelsestid i år[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: slut-lønn

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,700[a]	,490	,487	4893,33

a. Predictors: (Constant), Ansattes kjønn, Alder, Utdannelsestid i år

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,08E+10	3	3604216050	150,523	,000[a]
	Residual	1,13E+10	470	23944661,96		
	Total	2,21E+10	473			

a. Predictors: (Constant), Ansattes kjønn, Alder, Utdannelsestid i år

b. Dependent Variable: slut-lønn

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3952,764	1647,620		-2,399	,017
	Alder	17,526	19,917	,030	,880	,379
	Utdannelsestid i år	1378,187	86,967	,582	15,847	,000
	Ansattes kjønn	-3346,576	483,622	-,244	-6,920	,000

a. Dependent Variable: slut-lønn

Tabellen ”Modell summary ”viser at ”R^2 adjusted” har økt fra 0,438 til 0,490. 49% av variasjonen i sluttlønn kan forklares vha de tre uavhengige variablene alder, utdannelse og kjønn. 51% av variasjonen i sluttlønn er fortsatt uforklart. Studerer vi F-verdien og ”Sig.” ser vi at regresjonsmodellen samlet sett gir en signifikant forklaring av variasjonen i sluttlønn.

Men vi er opptatt av sammenhengen mellom én og én variabel og sluttlønn forutsatt at de andre variablene holdes konstant. Vi ser da på b-verdiene og tester nullhypotesene om at de virkelige betaene til alder, utdanning og kjønn er lik null. Siden "Sig." < 0,05[16] for utdannelse og kjønn kan vi på 5% signifikansnivå forkaste hyposten om ingen sammenheng. Alder er fortsatt ingen signifikant forklaringsvariabel. Men, vi beholder den inntil videre.

Legg for øvrig merke til at koeffisienten til kjønn er negativ. Dette kommer av at vi har satt variabelen kjønn =1 for kvinne. I denne stikkprøven tjener en kvinnelig ansatt 3346 mindre enn en mann, når vi holder utdanning og alder konstant. Men kanskje mennene har mer ansvarsfulle stillinger?

Samtidig ser vi at koeffisienten til utdannelse nå er redusert til 1378. Udanning gir ikke samme uttelling når vi samtidig justerer for effekten av kjønn. Dette kommer av at menn i denne stikkprøven har lengre utdanning enn kvinnene. Når vi samtidig nå vet at mennene tjener mer enn kvinner, når vi justerer for alder og utdanning, forstår vi at koeffisienten til utdanning i case 2 innholdt en skjult indirekte effekt fra utdanninglengde via kjønn til lønn.

6.2.5 Case 4

Vi foretar nå en regresjon på alle de 4 forklarende faktorene.

Vi må da gjøre et forarbeid. Vi har 7 yrkeskategorier og skal nå vha. menyvalget "Compute" lage 6 dummyvariable slik at:

$D_{kp} = 1\, kassepersonell$

$D_{kp} = 0\, ellers$

$D_{ka} = 1\, kasseaspirant$

$D_{ka} = 0\, ellers$

$D_{sv} = 1\, sik\,\text{ker}\, hetsvakt$

$D_{sv} = 0\, ellers$

$D_{kro} = 1\, kunderådg.\, under\ opplæring$

$D_{kro} = 0\, ellers$

$D_{kr} = 1\, kunderådg.$

$D_{kr} = 0\, ellers$

[16] Dette er ensidige tester; Beta utdanning >0 og Beta kjønn <0. Derfor sammenligner vi p-verdien med 5% ved et 5% signifansnivå.

$D_l = 1\, leder$

$D_l = 0\, ellers$

Den siste yrkesgruppen; tekniker, er her referansegruppen. Et kasus for en tekniker vil ha alle 6 dummyvariable lik 0.

For å få registrert disse variablene i vårt dataark, kan vi bruke kommandoen ”Recode”. Aller først har vi vært inne i delvinduet ”Variable View” og sjekket hvordan yrkesgruppene er nummerert under variabelen yrke. Vi fant:

1: kassepersonell
2: kasseaspirant
3: Sikkerhetsvakt
4: Kunderådgiver under opplæring
5: Kunderådgiver
6: Leder
7: Tekniker

Vi velger fra menyen:

Transform
Recode
Into different variables

Vi kommer da inn i ”Recode different variables” arket. Vi velger variabelen ”Yrke” inn i ”Input Variables” boksen. Vi beveger oss så over i ”Output Variables” boksen. Her legger vi inn navnet på den første dummyvariabelen over; ”Dkp”, og klikker på ”Change”. Vi får da det nye variabelnavnet flyttet opp i boksen sammen med ”Yrke”. Det kan være lurt å legge til en forklarende tekst under ”Label” for eksempel; ”Dummyvariabel for kassepersonell”. Vi trykker så på ”Old and New Variables”-valget for å angi hvilke verdier på den gamle variabelen som skal knyttes opp mot bestemte verdier i den nye variabelen. I den nye dialogboksen har vi gamle verdier for Yrke (1-7) til venstre og de nye knyttet til dummyvariabelen ”Dkp” (0 eller 1) til høyre. Vi skriver inn tallet 1 under ”Old Value”, beveger oss til høyre og legger inn 1 for ”Dkp”. Vi trykker på ”Add”. Deretter tilbake til venstre side og krysser av på ”All Other Values” som blir kodet lik 0 under ”New Value”. Vi trykker igjen ”Add”. Vi har nå at alle kasus med kassepersonell har en ”Dkp” verdi lik 1, mens andre kasus har en ”Dkp” verdi lik 0.

Vi fortsetter på samme måte for hver Dummyvariabel.

Sjekk til slutt at alle yrkeskategorier unntatt tekniker har kun én dummyvariabel med verdi lik 1. Siden tekniker yrket er referansen, skal alle kasus der den ansatte er tekniker ha samtlige dummyvariabel for yrke lik null.

Vi kan nå kjøre en multippel regresjon der sluttlønn er forklart ved alder, utdanning, kjønn (én dummy) og yrke (6 dummies). Dette gir en regresjonsmodell med 9 uavhengige variable. Vi velger:

Analyze
Regression
Linear

Vi velger ”Sluttlønn” som ”Dependent Variable” og ”Utdannelsestid i år”, ”Alder”, ”Ansattes kjønn”, ”dkp”, ”dka”, ”dsv”, ”dkro”, ”dkr” og ”dl” som ”Independent Variables”. Vi trykker OK, og får følgende resultat:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	DL, DSV, DKR, DKRO, DKA, Ansattes kjønn, Alder, Utdannelsestid i år, DKP[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: slut-lønn

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,872[a]	,761	,756	3374,78

a. Predictors: (Constant), DL, DSV, DKR, DKRO, DKA, Ansattes kjønn, Alder, Utdannelsestid i år, DKP

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,68E+10	9	1864676289	163,724	,000[a]
	Residual	5,28E+09	464	11389122,12		
	Total	2,21E+10	473			

a. Predictors: (Constant), DL, DSV, DKR, DKRO, DKA, Ansattes kjønn, Alder, Utdannelsestid i år, DKP

b. Dependent Variable: slut-lønn

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	29051,465	2157,929		13,463	,000
	Utdannelsestid i år	527,152	75,787	,223	6,956	,000
	Alder	-44,232	16,187	-,076	-2,732	,007
	Ansattes kjønn	-1968,597	368,330	-,144	-5,345	,000
	DKP	-21832,3	1455,506	-1,598	-15,000	,000
	DKA	-22185,2	1496,585	-1,471	-14,824	,000
	DSV	-19739,8	1635,216	-,671	-12,072	,000
	DKRO	-12380,8	1494,594	-,510	-8,284	,000
	DKR	-10750,6	1506,333	-,395	-7,137	,000
	DL	-10597,9	2053,663	-,159	-5,160	,000

a. Dependent Variable: slut-lønn

Dette var et spennende resultat! Vi ser at forklaringskraften til modellen; ”R square”, har økt fra 0,49 til hele 0,761. 76,1% av variasjonen i sluttlønnen er forklart av denne modellen.

Og ikke nok med det,- samtlige forklaringsvariable er signifikante. Ved å inkludere et gjennomtenkt sett med forklaringsvariable kan vi nå og tolke beta-verdiene på en mer direkte måte. Beta for utdannelsestid er nå effekten av et år mer utdannelse på sluttlønnen når vi justerer for effekten av alder, kjønn og ulik yrkesgruppe.

Koeffisientene til alle 6 dummyvariable for yrke er negativ. Dette er fordi yrkeskategorien ”Tekniker” er referanse. Teknikerne har den høyeste gjennomsnittslønn. Koeffisienten forann ”Dka” har den største negative verdien. En kasseaspirant vil i denne stikkprøven ha en gjennomsnittslønn som er hele 22.185 mindre enn en tekniker. Husk at koeffisienten må tolkes som effekten av yrkeskategori når vi holder de andre uavhengige variablene i modellen fast. Vi får derfor et litt annet resultat enn hvis du sjekker forskjellene i gjennomsnittslønn for yrkesgruppene (vha ”Compare Means” for eksempel).

Koeffisienten foran kjønn er nå på hele -1.968. I stikkprøven tjener en kvinne i gjennomsnitt 1.968 mindre enn en mann når vi justerer for forskjeller i stillingstype, alder og utdanning. Det er en betydelig forskjellsbehandling. Det kan selvsagt være at modellen ikke godt nok fanger opp forhold som personlige egenskaper, arbeidskapasitet etc.. Det er likevel grunn til å være skeptisk til bransjens lønnspolitikk.

Alder har igjen fått en koeffisient med et fortegn motsatt av det vi forventet. En person som er ett år eldre enn en annen vil i denne stikkprøven i gjennomsnitt tjene 44 mindre når vi holder faktorer som utdanning, kjønn

og yrkeskategori konstant. Dette er et merkelig resultat. Det kan skyldes flere forhold:

1. Variabelen ”Alder” måler relevant yrkeserfaring på en dårlig måte. Jeg har forsøkt å erstatte variabelen alder med variabelen ”År” i tilsvarende stilling. Resultatet blir om lag likt.

2. Det finnes andre forklaringsvariable som ikke er fanget opp i modellen. Disse er korrelert med alder. Eksempler på dette er; omstillingsevne, arbeidskapasitet, teknologisk innsikt. Det kan og være at det er en samvariasjon mellom hudfarge og alder som slår negativt ut.

3. Det er ingen sammenheng mellom lengde på yrkeserfaring og lønn. Variabelen bør tas ut av modellen.

Det er mulig at vi burde kutte ut variabelen pga dårlig validitet. Jeg vil likevel råde til at vi forsøkte å spesifisere modellen enda bedre, før vi konkluderte endelig.

Én måte å vurdere hvorvidt modellen er godt spesifisert får vi ved å studere residualene. Det skal vi gjøre i neste avsnitt.

6.3 Analyse av residualene

I eksempelet over vil vi for hver av de 474 kasusene ha en observert verdi Y for sluttlønn. Avviket mellom den faktiske sluttlønnen og modellens predikerte sluttlønn gitt kjønn, alder, utdanning og yrkeskategori er lik residualene. Dette er den uforklarte delen av variasjonen i sluttlønn. I den siste regresjonsmodellen over, er denne uforklarte andelen av variasjonen til sluttlønnen redusert til 24%.

For å vurdere om modellen er veldefinert må vi sjekke at variansen til residualene er rimelig lik for ulike nivå av forklaringsvariablene Xi. Vi må og vurdere om mønsteret i residualene tyder på at de er uavhengig av hverandre; dvs. ikke tenderer til å ligge over eller under regresjonsplanet for ulike intervall av Xi.

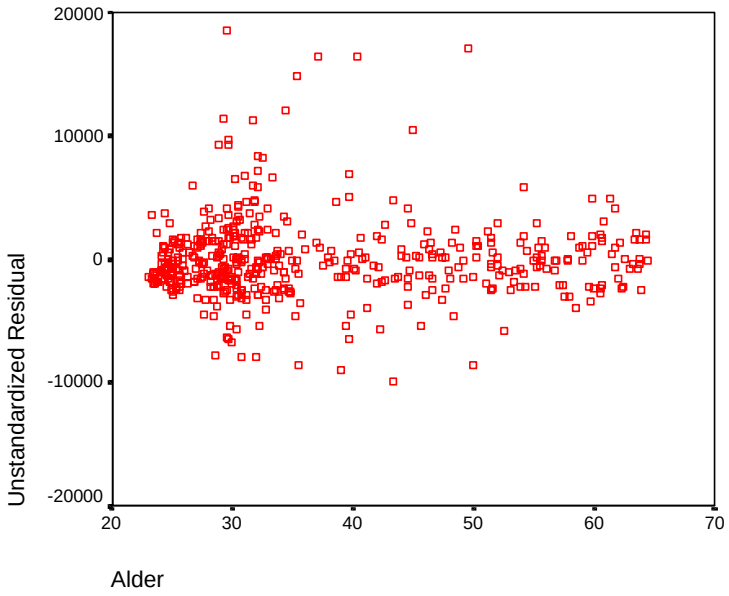
Vi velger fra menyen:

Analyze
    Regression
        Linear

Velg ”Slutløn” som ”Dependent Variable” og ”Alder”, ”Kjønn”, ”Utdanning”, ”Dkp”, ”Dka”, ”Dsv”, ”Dkro”, ”Dkr” og ”Dl” som ”Independent Variables”. Trykk på "Save". Her kan du velge å ta vare på noen nye variable som regresjonsanalysen lager. Du vil finne de igjen i dataarket. Under ”Predicted values” velger vi ”Unstandardized”. Vi får da

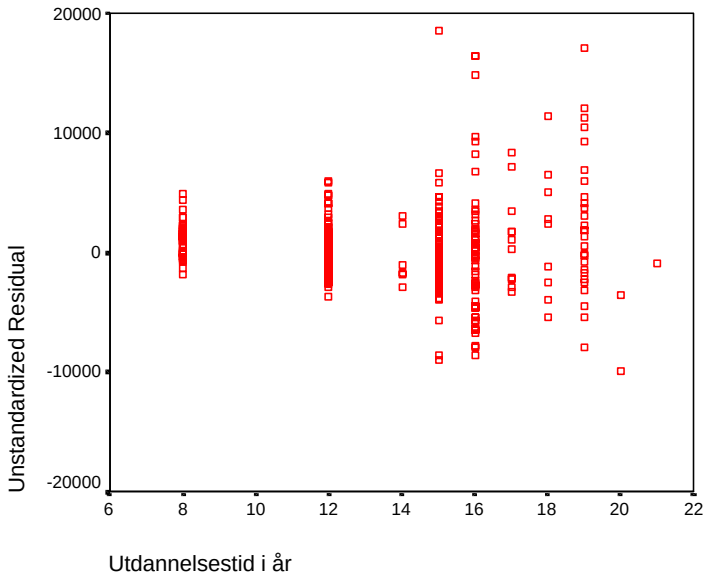
predikert Y verdi, eller $\hat{Y}_i$, i regresjonsmodellen. Under "Residuals" velger vi også "Unstandardized". Vi får da avviket mellom predikerte og faktiske Y-verdier; $\hat{Y}_i - Y_i$. Vi trykker OK, men bryr oss ikke om det resulterende utskriften.

I stedet går vi tilbake og velger "Graphs" i menyen og videre "Scatter" og "Simple" under "Define". Vi setter "Unstandardized residuals" som Y og "Alder" som X og får følgende graf:



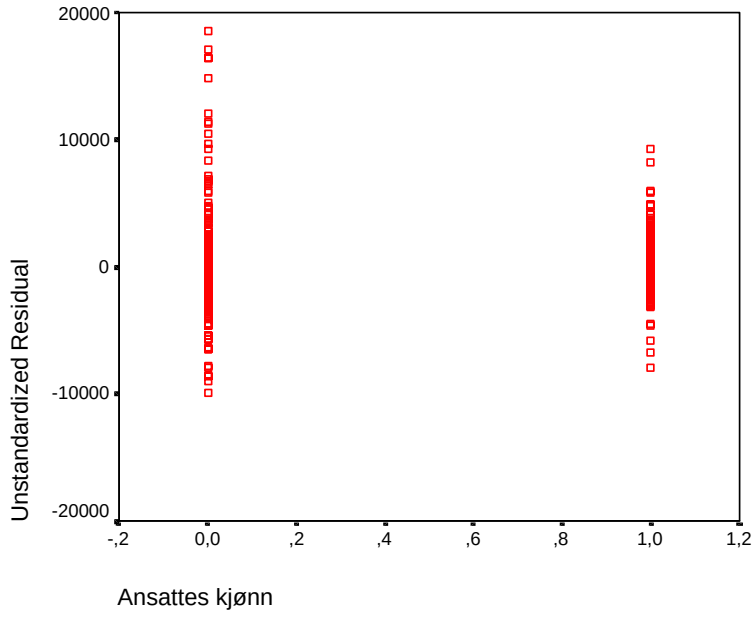
Residualene er spredt rundt forventingsverdien null på en tilsynelatende tilfeldig måte. Jeg synes imidlertid å kunne lese av plottet at spredningen i lønn for høye aldersgrupper er mindre enn for de unge. En forklaring på dette kan være at vi har utelukket forhold i modellen som tillegges vekt ved avlønning, og at disse påvirker spesielt lønnsnivået til de bankansatte mellom 30 og 45.

Vi lager nå tilsvarende skatterplott for residualene mot utdanningslengde.



Vi ser her tydelige tegn på at variansen til residualene øker med økt utdannelse. D.v.s. at vi også her har heteroskedastisitet i residualene. De største residualene bør her undersøkes ved at vi går tilbake til datagrunnlaget og spørreskjemane og sjekker at tallene er lagt inn korrekt. Under forutsetning av at datagrunnlaget er korrekt, synes jeg å kunne ane en viss kurve i residualene oppover; dvs. at effekten av ett års utdannelse ikke er stabil, men øker eksponensielt. Dette kan vi analysere ved å formulere en eksponensiell vekst modell.

Til slutt tar vi med et eksempel på skatterplott av residualene mot en dummyvariabel; her kjønn.



Skalaen er litt misvisende her. Kjønn er enten lik 1 (kvinne) eller 0 (mann). Vi ser at spredningen i lønn er større for mennene enn for kvinnene i

stikkprøven. Det er mulig at en bedre spesifisert modell, som inkluderte nye forklaringsvariable, ville gitt en større grad av samsvar mellom spredningen til residualene til kvinnene og mennene.

6.4 Ikke-lineære sammenhenger

6.4.1 Lineær transformasjon; eksponensiell vekst

Analysen av de residuale plottene i avsnittet over, tyder på en mulig eksponsiell sammenheng mellom utdannelse og lønn.

Den positive effekten på lønn målt i kroner av å ha ett år mer utdannelse er sterkere når du tar en mastergrad enn når du nettopp har startet opp et treårig bachelorstudium. Dette høres rimelig ut - i hvertfall når du analyserer samme type bransje eller samme type arbeid. Jeg er likevel litt usikker i utgangspunktet. Siden variabelen utdanningtid i år ikke skiller mellom relevant og ikke relevant utdanning kan vi få en beta som ikke er signifikant. Det er og mulig at det finnes et optimalt utdanningsnivå, at ytterligere utdanning i ekstreme tilfeller kan oppfattes som negativt. Siden jeg ikke har en klar á priori oppfatning, vil signifikanssannsynligheten til beta avgjøre hvorvidt jeg forkaster nullhypotesen i følgende modell:

$$Y = Ae^{bX_{utd}}$$
$$\ln Y = \ln A + bX_{utd}$$
$$\ln Y = a + bX_{utd}$$

$$H_1 : \beta > 0$$
$$H_0 : \beta = 0$$

For å lage denne regresjonsmodellen, må vi først definere en ny variabel; ln sluttlønn. Dette gjør vi vha menyvalget:

Transform
    Compute

Vi kommer inn i ”Compute variable” arket. Under ”Target variabel” skriver vi navnet på den nye variabelen ”Lnsllønn”. Under ”Type and labels” legger vi inn en forklarende tekst; ”LN til sluttlønn”. Vi går deretter ned og velger ”Lnnumexpr()” under ”Functions”. Vi bruker pil opp for å få den opp i ”Numeric expression” boksen. Deretter går vi ned i variabellisten og piler inn sluttlønn i parantesen til funksjonen. Vi trykker OK. Tilbake i dataarket finner vi nå en ny variabel i siste kolonne; ”Lnsllønn”.

Nå er vi klar til å utføre regresjonen mellom ”Lnslønn” og den uavhengige variabelen ”Utdanningstid”. Vi velger:

Analyze

Regression

Linear

Vi velger ”Lnsllønn” som ”Dependent Variable” og ”Utdanningstid” som ”Independent variable”.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Utdannelsestid i år[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: ln til sluttlønn

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,697[a]	,485	,484	,2853

a. Predictors: (Constant), Utdannelsestid i år

b. Dependent Variable: ln til sluttlønn

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36,251	1	36,251	445,300	,000[a]
	Residual	38,424	472	8,141E-02		
	Total	74,675	473			

a. Predictors: (Constant), Utdannelsestid i år

b. Dependent Variable: ln til sluttlønn

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8,146	,063		129,839	,00000000000000000
	Utdannelsestid i år	9,596E-02	,005	,697	21,102	,00000000000000000

a. Dependent Variable: ln til sluttlønn

Casewise Diagnostics[a]

Case Number	Std. Residual	ln til sluttlønn
432	3,329	10,63
456	3,337	10,63
461	3,208	10,60
471	3,251	10,90

a. Dependent Variable: ln til sluttlønn

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	8,9135	10,1610	9,4405	,2768	474
Residual	-,6626	,9522	-4,79E-15	,2850	474
Std. Predicted Value	-1,904	2,603	,000	1,000	474
Std. Residual	-2,322	3,337	,000	,999	474

a. Dependent Variable: ln til sluttlønn

Dette var interessant. Koeffisientene i modellen er helt klart signifikante og vi kan forkaste nullhypotesen om at Beta til utdannelsestid i den eksponensielle funksjonen over er lik null. Vi setter regresjonens koeffisienter inn i formelen og får:

$$\ln Y = a + bX_{utd}$$
$$\ln Y = 8{,}146 + 0{,}09596 X_{utd}$$
$$Y = e^{8{,}146} e^{0{,}09596 X_{utd}}$$
$$Y = 3450 e^{0{,}09596 X_{utd}}$$

Dette tolker vi som at en endring i utdannelse med ett år gir en %vis økning i lønn på tilnærmet 9,6%. Når vi har e som grunntall i denne multiplikative modellen vil b være tilnærmet lik den %vise endringen i Y ved en enhets endring i X. Jo mindre endringen i X er, jo bedre blir tilnærmingen. For eksempel gir modellen følgende effekt av 3 og 4 års utdannelse:

$$Y(4) = 3450 e^{0{,}09596*4} = 5064$$
$$Y(3) = 3450 e^{0{,}09596*3} = 4601$$
$$\frac{(Y(4) - Y(3))}{Y(3)} = \frac{5064 - 4601}{4601} = 0{,}10$$

6.4.2 Lineær transformasjon; polynomer

Vi henter eksempelet fra læreboka til Wonnacott & Wonnacott; tabell 14-2 side 450. Merk dere at dette er et styrt eksperiment, og at vi derfor kan se bort fra indirekte effekter.

Vi antar at sammenhengen mellom Avling (Y) og gjødsel (X) følger en annengradsligning der sammehengen først er positiv for så og snu og bli negativ:

$$\hat{Y} = b_0 + b_1 X + b_2 X^2$$
$$X_1 \equiv X$$
$$X_2 \equiv X$$
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Våre to hypoteser formulerer vi slik:

$$H_0 : \beta_{X1} = 0$$
$$H_1 : \beta_{X1} > 0$$

$$H_0 : \beta_{X2} = 0$$
$$H_1 : \beta_{X2} < 0$$

Merk at vi antar at fortegnet foran X^2 er negativt, mens fortegnet foran X er positivt. Dette gir en annengradsligning som stiger for lave verdier av gjødsel, mens den etter å ha passert ett kritisk nivå vil synke.

Dataene vi skal bruke er slik:

Avling	X1: Gjødsel	X2: Gjødsel^2
55	1	1
70	2	4
75	3	9
65	4	16
60	5	25

Vi henter fram et tomt dataark i ”Data Editor”. Først går vi til ”Variabel View” og legger inn egenskapene til de to første variablene; ”Avling” og ”Gjødsel”. Dette er numeriske variable og vi skal ikke ha med desimaler. Antall tegn kan reduseres til 2 og 1.

Deretter går vi inn i ”Data View” arket og skriver inn de 5 observasjonene til hver av de to variablene.

Vi mangler en variabel; X2 eller gjødsel kvadrert. Siden denne variabelen er en matematisk utregning med utgangspunkt i variabelen X1: gjødsel, så bruker vi funksjonen ”Compute”. Velg fra menyen:

Transform
    Compute

I dialogboksen går vi opp i ”Target variables” og kaller den nye variabelen ”Gjkvadr”. Vi gir den et mer utfyllende navn under ”Type and Labels”. Jeg

piler ”Gjødsel” opp i ”Numeric Expression” boksen. Trykker så på **, som jeg vha høyre musetast har funnet ut betyr opphøyd i,. og trykker så på 2. Den nye variabelen skal nå bli gjødsel**2 eller gjødsel^2. Jeg trykker OK, og sjekker i dataarket at vi har fått en ny variabel med de ønskede tall.

Resten er lett. Vi velger fra menyen:

Analyze
Regression
Linear

Vi velger ”Avling” som ”Dependent Variabel” og ”Gjødsel” og ”Gjkvadr” som de to uavhengige variablene og trykker OK.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Gjødsel kvadrert, GJØDSEL[a]	,	Enter

a. All requested variables entered.

b. Dependent Variable: AVLING

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,935[a]	,874	,749	3,96

a. Predictors: (Constant), Gjødsel kvadrert, GJØDSEL

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	218,571	2	109,286	6,955	,126[a]
	Residual	31,429	2	15,714		
	Total	250,000	4			

a. Predictors: (Constant), Gjødsel kvadrert, GJØDSEL

b. Dependent Variable: AVLING

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	36,000	8,502		4,234	,052
	GJØDSEL	24,071	6,479	4,814	3,715	,065
	Gjødsel kvadrert	-3,929	1,059	-4,805	-3,708	,066

a. Dependent Variable: AVLING

Forklaringskraften til modellen er sterk; R^2 = 0,874. Om lag 87% av variasjonen i avlingsnivå er her forklart vha ulik gjødselsmengde. Koeffisientene har rett fortegn. T-verdiene, som vi kan regne selv ved å dele b/SE, er store. Men, fordi stikkprøven er så liten blir likevel p-verdiene litt større enn 5%.

Skal du sjekke dette resultatet slår du opp i en tabell for t-fordelingen. Antall frihetsgrader = 5-3=2 fordi vi bruker 3 frihetsgrader til å estimere b0, b1 og b2. Kritisk t-verdi for en ensidig test når vi krever 5% signifikansnivå, er 2,92. På denne bakgrunn forkaster vi de to nullhypotesene formulert over.

I tabellene til SPSS fører de opp som standard p-verdien for tosidige tester. Da blir kritisk verdi - 4,30 og +4,30. Vi ser at ingen av de to uavhengige variablene har t-verdier som ligger utenfor disse kritiske verdiene. P-verdien i tabellen (=Sig.) reflekterer som standard en to sidig test.

6.4.3 Lineær transformasjon; konstant elastisitet

Eksempelet nedenfor er basert på en eksamensoppgave i faget BD616, våren 2004.

Personlig forbruk i et samfunn kan deles inn i utgifter til tjenester (for eksempel frisørbesøk), utgifter til varige forbruksgoder (for eksempel bilkjøp) og utgifter til ikke varige konsumgoder (for eksempel mat).:

Jeg har antatt følgende ikke lineære sammenheng mellom totalt forbruk; X, og utgifter til varige forbruksgoder;Y:

$$Y = A \cdot X^{\beta} \cdot \tau$$

En lineær transformasjon av denne modellen er gitt ved:

$$\ln Y = \alpha + \beta \ln X + \varepsilon$$
$$der: \alpha = \ln A, \varepsilon = \ln \tau$$

Utgiftstype:	Variabelnavn:	Måleenhet:
utgifter til tjenester	”EXPSERVICES”	Mrd. dollar faste 92-priser
utgifter til varige konsumgoder	”EXPDUR”	Mrd. dollar faste 92-priser
utgifter til ikke varige konsumgoder	”EXPNONDUR”	Mrd. dollar faste 92-priser
Totalt personlig forbruk	”PCEXP”	Mrd. dollar faste 92-priser

En estimering av denne modellen kan utføres basert på amerikanske data for perioden 1. kvartal 1993 til 3.kvartal 1998. Tallene er i milliard dollar i faste 1992 priser. Du finner datagrunnlaget i filen: **”Gujarati tab 6 3.sav”.**

Før vi estimerer regresjonsmodellen, må vi regne ut to nye variabler:

lnX: den naturlige logartimen til totaltforbruk; LNPCEXP og
lnY: den naturlige logaritmen til utg. til varige forbr.goder; LNEXPDUR

Dette gjør vi vha menyvalget:

Transform
Compute

Bruk framgangsmåten skissert i avsnittet ”Lineær transformasjon; eksponensiell vekst”. Gi gjerne variablene et mer forståelig navn under ”Label” i ”Variabel View” arket.

Nå er vi klar til å utføre regresjonen mellom ”LNEXPDUR” og ”LNPCEXP”. Vi velger:

Analyze
Regression
Linear

Vi velger ”LNEXPDUR” som ”Dependent Variable” og ”LNPXEXP” som ”Independent variable”. Resultatet av estimeringen er gitt under:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	logartimen til totalt personlig forbruk[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: logartimen til forbruk på varige konsumgoder

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,992[a]	,985	,984	,01332

a. Predictors: (Constant), logartimen til totalt personlig forbruk

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,244	1	,244	1376,130	,000[a]
	Residual	,004	21	,000		
	Total	,248	22			

a. Predictors: (Constant), logartimen til totalt personlig forbruk

b. Dependent Variable: logartimen til forbruk på varige konsumgoder

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9,697	,434		-22,337	,000
	logartimen til totalt personlig forbruk	1,906	,051	,992	37,096	,000

a. Dependent Variable: logartimen til forbruk på varige konsumgoder

Til eksamen våren 2004 ble studentene stilt følgende spørsmål:

a) Sett opp den estimerte modellen
b) Hvordan tolker du den estimerte betakoeffisienten
c) Sett opp et 95% konfidensintervall for beta.
d) Forklar hvordan du kan bruke konfidensintervallet over ved hypotesetesting av beta. Gi et eksempel.

Basert på utskriften over, skal du nå være i stand til å gi følgende svar:

a) Basert på kolonne B i tabellen "Coefficients" kan du sette opp modellen slik:

$$\ln \hat{Y} = -9{,}697 + 1{,}906 \ln X$$

b) $$\beta \approx \frac{\partial Y / Y}{\partial X / X}$$

Betaen er en elasitistet som viser hvor forholdet mellom prosentvis endring i Y og prosentvis endring i X. Her har vi at en oppgang i X med 1% gir en oppgang i Y med 1,906*1%=1,906%. Utgifter til

varige forbruksgoder er m.a.o. sterkt prosykliske og vil stige (synke) mer enn andre deler av forbruket i perioder med positiv vekst.

c) For å sette opp konfidensintervallet trenger du opplysninger fra utskriften; Betaestimatet og standard feil til dette estimatet. I tillegg må du slå opp i en tabell over t-verdier. t-verdien skal reflektere valgt signifikansnivå (for eksempel 5% totalt) og antall frihetsgrader; antall observasjoner – antall estimater (her: 24-2 = 22).

$$\hat{\beta} \pm t^{22}_{0,025} \cdot SE$$
$$1{,}906 \pm 2{,}074 \cdot 0{,}051$$
$$\{1{,}800 - 2{,}012\}$$

d) Konfidensintervallet innholder alle akseptable nullhypoteser ved 5% signifikansnivå.

Et eksempel på bruk av konfidensintervallet er følgende hypotese:

$$H_0 : \beta = 1$$
$$H_1 : \beta \neq 1$$

Siden konfidensintervallet ikke inkluderer 1, forkaster vi nullhypotesen.

7 Diskriminantanalyse

Diskriminantanalyse er en teknikk som er relevant å bruke når den avhengige variabelen er en kategorisk variabel og de uavhengige variablene er metriske (intervall eller forholdstall). I noen tilfeller har den avhengige variabelen to kategorier, for eksempel brukere kontra ikke brukere av et produkt, menn kontra kvinner osv. I andre tilfeller er det snakk om flere enn to kategorier, som for eksempel god kunde, middels god kunde og dårlig kunde. Dersom det er snakk om to kategorier, kalles teknikken to-gruppe-diskriminantanalyse. Når det er snakk om tre eller flere kategorier, kalles teknikken multippel diskriminantanalyse.

Diskriminantanalyse er en teknikk som er mye brukt i markedsanalyser, og da særlig med tanke på segmentering. Her følger et intuitivt eksempel, hentet fra en undersøkelse av feriegjester gjennomført ved Norges Handelshøyskole: Analysen forsøkte å beskrive hva som skiller besøkende til Voss, Vesterålen og Stryn med henblikk på hvilke motiver de hadde for å feriere. Resultatet viste at de som besøkte Voss, la betydelig mindre vekt på at barna skulle ha et aktivitetstilbud og muligheter for fottur. Besøkende til Voss var opptatt av natteliv, underholdning, mulighet for kulturell opplevelse og muligheten til å kunne spise godt.

Besøkende til Stryn vart mest opptatt av aktivitetstilbud for barn og gode muligheter for fotturer. De var mindre opptatt av kulturelle opplevelser, natteliv og underholdning og muligheten for god mat. De samme karakteristikkene fant man for besøkende til Vesterålen.

Diskriminantanalysen viste dermed at gjestene på Voss hadde en mer ”urban” profil enn på de andre to stedene. Man fant dermed hva som *skiller* de tre stedene. Det som var felles på de tre stedene var mulighet for naturopplevelse.

7.1 To-gruppe-diskriminantanalyse

For å illustrere tilfellet med to-gruppe-diskriminantanalyse, bruker vi data fra en undersøkelse der respondentene svarer på om de kan tenke seg å motta nyheter interaktivt (via kabel) til sitt eget TV-apparat. Det interessante er å finne de segmentene som er interessert i et slikt tilbud. Flere demografiske variabler er kartlagt, som utdanning, alder, inntekt, kjønn, antall barn, antall organisasjoner man er med i og antall timer respondenten bruker foran TV-en hver dag. I tillegg er det en variabel som sier noe om respondenten aksepterer tilbudet eller ikke. Disse variablene kan vi kalle for $X_1 ..... X_2$. Modellen kan da settes opp slik:

$$D = b_0 + b_1 X_1 + b_2 X_2 + ... + b_7 X_7$$

I denne modellen vil da *D* være diskriminantscoren. Denne kan brukes til prediksjon. Dersom vi her for eksempel ser på en potensiell bankkunde, kan man legge inn variabelverdiene for denne kunden og koeffisientene i modellen. Diskriminantscoren vil da enkelt kunne regnes ut og om denne er høy eller lav vil kunne gi en prediksjon på om man her forholder seg til en potensiell god eller dårlig bankkunde. De estimerte koeffisientene i modellen er gitt ved $b_0.....b_7$. Her vil b_0 være det estimerte konstantleddet, mens b_1 sier noe om hvor godt den første variabelen X_1, utdanning, skiller tilfellene. En høy verdi på b_1 betyr da at utdanning bidrar mye til å skille mellom gode og dårlige bankkunder. Dersom derimot b_1 er lav eller nær 0, vil dette si at utdanning ikke betyr mye for om en kunde er god eller dårlig.

Vi ønsker nå å se nærmere på markedsundersøkelsen om interaktive nyheter. Målet er å sette opp en modell som predikerer om en kunde vil komme til å kjøpe produktet eller ikke. I denne undersøkelsen blir det også spurt om kunden kan tenke seg å kjøpe produktet eller ikke. På denne måten får vi en "fasit" der vi kan sjekke hvor god modellen vår blir. Dette kommer vi tilbake til senere. En slik "fasit" vil man ikke ha når man skal bruke modellen på nye data, for eksempel på en ny potensiell kunde. "Fasiten" illustrerer videre nytten av bruke en diskriminantmodell i forhold til ikke å ha en modell i det hele tatt.

Vi henter fila **nyhetskanal.sav**:

File
    Open
        Data

Oppgave: Finn ut hva som er gjennomsnittlig utdanning, alder, timer med TV-kikking daglig og inntekt hos respondentene.

Etter den innledende inspeksjonen kjører vi nå diskriminantanalysen:

Analyze
    Classify
        Discriminant

I dialogboksen setter vi variabelen "Nyheter" inn i Grouping Variable boksen. Legg merke til de to spørsmålstegnene som nå følger etter variabelen. Disse kommer frem fordi diskriminantanalyse også kan brukes i tilfeller der den avhengige variabelen har flere enn to kategorier. Programmet trenger derfor at vi spesifiserer maksimums- og minimumsverdien på den avhengige variabelen. Dette gjøres ved å trykke på Define Range i dialogboksen.

Define Range

Skriv 0 i minimumsfeltet og 1 i maksimumsfeltet.

Continue

Lenger nede i dialogboksen ligger feltet der man spesifiserer hvilke variabler som skal med som uavhengige variabler. Dersom man ikke har noen konkret modell i tankene, kan det være fornuftig å ta med alle, og så velge stepwise som metode. Noen foretrekker denne metoden for da kan man, ifølge SPSS, etter hvert ta bort de variablene som betyr minst i estimeringen. Har man med mange variabler vil det imidlertid kunne være et problem med multikolinearitet. Dette vil si at de uavhengige variablene er høyt korrelert med hverandre og man kan da få problem med at estimatene er upresise fordi standardavviket øker. Dette skaper problem for statistisk testing av estimatene.

Flytt alle variablene over til boksen independents. Velg så:

Use stepwise method
    Method

Man får nå frem en dialogboks som spesifiserer hvilken metode som skal benyttes. Standardvalget her er Wilk's Lambda, en metode som også er den mest benyttede. Vi velger derfor å holde oss til denne. Programmet kommer nå med et forslag til hvilke kritiske F-verdier vi bør velge mht å ta med variabler i estimeringene. Disse kan justeres dersom man ønsker det, eventuelt justeres gjennom sannsynlighets anslag. Det typiske vil være at man setter opp et strengere kriterium dersom man har et stort utvalg. Gjør nå følgende:

Continue
    Classify
        Summary table

I den dialogboksen som nå kommer frem, Classification dialog-boks, kan man spesifisere hva som skal vises i utskriften. Ved å bestille en Summary table, kan man effektivt evaluere diskriminantanalysen som er gjort. Vi sjekker videre at alle standard forslagene fra programmet er med. Dette er "All groups equal" (under Prior Probabilities), som sier noe om at en observasjon med like stor sannsynlighet hører til hver av kategoriene. Dersom man på forhånd vet noe annet, kan dette spesifiseres. Dette er imidlertid sjelden nødvendig. Videre ser man om "Within-groups" (under use Covariance matrix) er valgt. Vi gjør oss nå ferdige med Classification dialog boksen og spesifiserer hvilke statistiske rapporter man ønsker å bestille:

Continue
    Statistics
        Means
        Univariate ANOVAs
        Fisher's
        Unstandarized

Fisher's-koeffisientene og de ustandardiserte diskriminantkoeffisientene kan brukes til å predikere på fremtidige observasjoner. Dersom det bare er to mulige kategorier på den avhengige variabelen, slik det er i vårt tilfelle der denne variabelen enten har verdien "kjøp" eller "ikke kjøp" nyhetskanal, vil begge både være fornuftige og enkle å bruke. Dersom man kjører diskriminantanalyse med flere enn to ufallskategorier, vil Fisher's-koeffisientene være de mest fornuftige å bruke.

Dersom man ønsker å se på korrelasjoner mellom gruppene i de uavhengige variablene kan man velge within-groups correlations. Dersom man heller er interessert i å se på kovariansen mellom de uavhengige variablene, kan man bestille en eller flere av de tre mulige kovariansmatrisene. Dette er ikke alltid nødvendig. Det mest nyttige er å bestille korrelasjonsmatrisen med tanke på å undersøke om multikolinearitet er et problem. Dersom variablene er svært høyt korrelert, kan dette gi upresise estimat på koeffisientene på grunn av høy grad av multikolinearitet.

Vi avslutter nå statistikkdialogboksen og går videre:

Continue

Dersom man ønsker å lagre variabler etter analysen, kan man gå inn på Save. Her er det en mulighet for å lagre opp til tre variabler. En av disse er "predicted group membership", som er en variabel som forteller oss hvilke gruppe tilfellene faller inn i. For eksempel vil et individ som får verdien 0 på denne variabelen være predikert til å svare at vedkommende ikke ønsker å kjøpe nyhetskanal. Tilsvarende vil et individ som får verdien 1 bli predikert til svarkategori "ønsker å kjøpe nyhetskanal". De to andre variablene man kan velge å lagre er "Discriminant scores" og "Probabilies for group membership". Den første av disse er diskriminant scoren, mens den andre er sannsynligheten for å få verdien 0 på variabelen "predicted group membership". Med andre ord, sannsynligheten for å bli predikert til kategorien "ikke ønsker å kjøpe nyhetskanal".

Vi avslutter nå diskriminantanalyse dialogboksen og kjører analysen. Trykk derfor:

Ok

I output-vinduet ser vi nå resultatene etter analysen. Det første man gjør, er å dra i heisen til man kommer ned til den tabellen der man klassifiserer resultatene. Dette er en viktig tabell siden den forteller oss hvor godt modellen vår predikerer. I radene ser vi hva respondentene faktisk har svart, mens i kolonnene ser vi hva modellen predikerer. Vi tester med andre ord modellen mot "fasiten". Dette kan vi gjøre når vi har historiske data eller, som i vårt tilfelle, vet om kundene kommer til å kjøpe produktet eller ikke. Her ser man at 227 av respondentene svarer at de ikke kan tenke seg å kjøpe nyhetskanal, mens 214 kan tenke seg dette. Av de 227 som svarer at de ikke kan tenke seg å handle, plasserer modellen 157 riktig (69,2 %) og 70 (30,8 %) feil. For de 214 som sier de kan tenke seg å handle, plasserer modellen

142 riktig (66,4 %) og 72 feil (35,6 %). Ser vi hele modellen under ett, er 67,8 % av respondentene plassert riktig i modellen.

Spørsmålet blir da om dette er et bra resultat. En modell som ”bare ” predikerer 67,8 % av tilfellene riktig høres intuitivt sett kanskje noe dårlig ut. Man må imidlertid ha klart for seg hva man sammenligner med. Det kunne selvfølgelig vært ønskelig med 100 % treffsikkerhet, men hva om man istedenfor gjettet at alle hører til i gruppen som ikke ønsker å kjøpe nyhetskanal? Da ville vi hatt rett i 227 av 441 (227+214) tilfeller, eller 51,5 % av tilfellene. En modell som predikerer 67,3 % av tilfellene riktig er dermed atskillig bedre enn å gjette. Nå er dette eksempelet noe ”rart” i den forstand at man trenger jo ikke spørre om mange bakgrunnskjennetegn for å predikere om en person skal kjøpe produktet eller ikke. Man kan jo bare stille spørsmålet direkte og få det riktige svaret med en gang? Jo, i dette tilfellet kan man det. Men diskriminantanalyse brukes først og fremst i situasjoner der man ikke kan stille et sånt spørsmål direkte. Man kan for eksempel ikke spørre en potensiell bankkunde om han er en god eller dårlig betaler. Dersom man likevel spør får man helt sikker skeive svar. Man kan derimot godt spørre om bakgrunnskjennetegn som alder, lønn, sivil status osv. Ut fra disse opplysningene kan man så predikere om dette sannsynligvis blir en god eller dårlig bankkunde.

Classification Results(a)

		Aksepterer nyhetstilbudet	Predicted Group Membership		Total
			No	Yes	
Original	Count	No	158	69	227
		Yes	75	139	214
	%	No	69,6	30,4	100,0
		Yes	35,0	65,0	100,0

a 67,3% of original grouped cases correctly classified.

Vi er nå interessert i å finne ut hva som karakteriserer de som godtar tilbudet om å kjøpe nyhetskanal. Man kan nå enten dra i heisen til man kommer opp til tabellen Group Statistics, eller så kan man velge denne tabellen ut fra menyen til venstre i vinduet:

Å sammenligne gjennomsnittet i disse to gruppene gir i vårt tilfelle forholdsvis liten informasjon, bortsett fra når det gjelder variabelen alder. Her ser vi at gjennomsnittlig alder for de som svarer at de kan tenke seg å kjøpe det tilbudte produktet er i underkant av 43 år, mens for de som ikke vil kjøpe produktet er gjennomsnittsalderen ca 35,5 år. Gruppen som kan tenke seg å kjøpe produktet er altså rundt 7 år eldre i gjennomsnitt enn gruppen som ikke vil kjøpe.

Group Statistics

Aksepterer nyhetstilbudet		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
No	Antall år med utdanning	13,29	2,611	204	204,000
	Kjønn	,60	,492	204	204,000
	Alder i år	35,43	11,081	204	204,000
	Timer fremfor TV en hver dag	2,63	1,802	204	204,000
	Antall organisasjoner en er medlem av	1,54	1,617	204	204,000
	Antall barn	1,47	1,568	204	204,000
	Inntektskategori	2,97	1,641	204	204,000
Yes	Antall år med utdanning	13,74	2,739	189	189,000
	Kjønn	,45	,499	189	189,000
	Alder i år	42,85	12,191	189	189,000
	Timer fremfor TV en hver dag	2,57	1,651	189	189,000
	Antall organisasjoner en er medlem av	1,60	1,659	189	189,000
	Antall barn	1,79	1,623	189	189,000
	Inntektskategori	3,43	1,632	189	189,000
Total	Antall år med utdanning	13,51	2,679	393	393,000
	Kjønn	,53	,500	393	393,000
	Alder i år	38,99	12,192	393	393,000
	Timer fremfor TV en hver dag	2,60	1,729	393	393,000
	Antall organisasjoner en er medlem av	1,57	1,635	393	393,000
	Antall barn	1,63	1,600	393	393,000
	Inntektskategori	3,19	1,651	393	393,000

Vi ønsker nå å se på testen av gjennomsnittet mellom gruppene. Da drar vi i heisen til vi kommer til tabellen ”Tests of Equality of group Means”, eller så kan man velge denne tabellen ut fra menyen til venstre i vinduet:

Tests of Equality of Group Means

	Wilks' Lambda	F	Df1	df2	Sig.
Antall år med utdanning	,993	2,738	1	391	,099
Kjønn	,978	8,806	1	391	,003
Alder i år	,907	39,947	1	391	,000
Timer fremfor TV en hver dag	1,000	,122	1	391	,728
Antall organisasjoner en er medlem av	1,000	,126	1	391	,723
Antall barn	,990	4,029	1	391	,045

Inntektskategori	,980	8,029	1	391	,005

Testen vi bruker, Wilks' Lambda, vil for hver av de uavhengige variablene være forholdet mellom varians innenfor gruppen og total varians. Målet med diskriminantanalysen er å finne en funksjon som skiller gruppene godt, og som dermed gir liten varians innenfor gruppene og mye varians mellom gruppene. Målet er altså en liten verdi på Wilks' Lambda. Av tabellen over ser vi at variabelen alder har den laveste verdien på Wilks' Lambda og den høyeste F-verdien (og er dermed den mest signifikante variabelen). Alder er derfor den variabelen som først blir plukket ut i den stegvise analysen som programmet gjør. Dette kan vi se av tabellen under:

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	Alder i år	1,000	39,947	
2	Alder i år	,992	42,562	,978
	Kjønn	,992	11,304	,907
3	Alder i år	,971	46,237	,972
	Kjønn	,992	10,998	,893
	Antall år med utdanning	,978	5,920	,882

Alder blir altså først tatt ut i den stegvise analysen. Programmet gjør nå en ny test på de gjenværende variablene i datasettet. Dette kan sees under steg 1 i tabellen "Variables Not in the Analysis" som er vist under. Her ser vi at den variabelen som nå har høyest F-verdi og som dermed er mest signifikant, er kjønn. I tabellen over, "Variables in the Analysis", er begge variablene "Alder i år" og "kjønn" tatt med under steg 2. Programmet gjentar prosedyren med de gjenværende variablene og plukker nå ut "Antall år med utdanning" som variabel nummer 3. Steg 3 i tabellen under viser at det nå ikke er noen signifikante variabler igjen.

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	Antall år med utdanning	1,000	1,000	2,738	,993
	Kjønn	1,000	1,000	8,806	,978
	Alder i år	1,000	1,000	39,947	,907
	Timer fremfor TV en hver dag	1,000	1,000	,122	1,000
	Antall organisasjoner en er medlem av	1,000	1,000	,126	1,000
	Antall barn	1,000	1,000	4,029	,990
	Inntektskategori	1,000	1,000	8,029	,980
1	Antall år med utdanning	,978	,978	6,210	,893
	Kjønn	,992	,992	11,304	,882
	Timer fremfor TV en hver dag	,993	,993	,035	,907
	Antall organisasjoner en er medlem av	1,000	1,000	,092	,907
	Antall barn	,727	,727	2,087	,902
	Inntektskategori	,964	,964	2,506	,902
2	Antall år med utdanning	,978	,971	5,920	,869
	Timer fremfor TV en hver dag	,993	,985	,030	,882
	Antall organisasjoner en er medlem av	,997	,989	,012	,882
	Antall barn	,718	,718	1,103	,879
	Inntektskategori	,851	,851	,206	,881
3	Timer fremfor TV en hver dag	,909	,895	,830	,867
	Antall organisasjoner en er medlem av	,980	,962	,041	,868
	Antall barn	,681	,681	,259	,868
	Inntektskategori	,732	,732	,240	,868

Vi kan nå se på to overordnede vurderinger av diskriminantfunksjonen. Dette gjøres ut fra tabellene Eigenvalues og Wilks' Lambda. I tabellen Eigenvalues kan man se den kanoniske korrelasjonen, og som vi ser er denne 0,363 i vårt tilfelle. Den kanoniske korrelasjonen forteller noe om styrken i relasjonen mellom den avhengige variabelen (diskriminantscoren) og gruppene. I det tilfellet der vi har to grupper, vil den kanoniske korrelasjonen være det samme som Pearsons korrelasjonskoeffisient. En høy verdi på denne koeffisienten indikerer en god modell.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,151(a)	100,0	100,0	,363

a First 1 canonical discriminant functions were used in the analysis.

Wilks' Lamda i tabellen under viser andelen av den totale variansen i diskriminantscoren som ikke blir forklart av forskjell mellom gruppene. Av tabellen under ser vi at Wilks' Lambda er forholdsvis høy, 0,87 (den kan maks være 1). En Lambda på 1 oppstår dersom gjennomsnittlig diskriminantscore er lik i de to gruppene og når det ikke er noen variasjon mellom gruppene.

For at vi skal kunne bruke analysen, er det nødvendig at Lambda er signifikant. Til dette benyttes det en chi-kvadrat test. I vårt tilfelle ser vi at Wilks' Lambda er helt klart er statistisk signifikant, noe som indikerer at modellen er ok. Man skal likevel være forsiktig med å trekke videre konklusjoner. Det at modellen er signifikant betyr ikke nødvendigvis at den er effektiv. Vi har riktignok forkastet en nullhypotese om at gjennomsnittet er likt i de to gruppene. Små forskjeller i gjennomsnitt mellom gruppene kan være statistisk signifikante, men likevel gi en dårlig diskriminering mellom gruppene.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	,869	54,902	3	,000

Siden modellen sett under ett er signifikant og ser ut til å være ok, går vi nå over til å se på de enkelte koeffisientene. Tabellen "The standarized canonical discriminant function coeffisients" viser den relative viktigheten av de enkelte prediksjonsvariablene i diskriminantfunksjonen. Ikke overraskende er det i vårt tilfelle alder som er den dominerende prediksjonsvariabelen. Tolkningen her er den samme som vi ville hatt i tilfellet med multippel regresjon. At koeffisientene er standardisert betyr at de er standardisert til et gjennomsnitt på 0 og et standardavvik på 1. I dette tilfellet vil det ikke være noe konstantledd.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Antall år med utdanning	,341
Kjønn	-,459
Alder i år	,912

Canonical Discriminant Function Coefficients

	Function
	1
Antall år med utdanning	,128
Kjønn	-,927
Alder i år	,078
(Constant)	-4,296

Unstandardized coefficients

Tabellen ”canonical discriminant function coefficients” er egentlig den samme som tabellen foran, bortsett fra at koeffisientene ikke er standardisert. I dette tilfellet er det også med et konstantledd. De ustandardiserte koeffisientene sier heller ikke noe om den relative viktigheten av de ulike variablene. Disse koeffisientene kan imidlertid brukes til å plassere en observasjon på diskriminantfunksjonen som skiller gruppene. Dette vil si å predikere hvorvidt en observasjon eller individ kan kategoriseres i en av gruppene. Vi kan altså multiplisere verdiene til en observasjon med de ustandardiserte koeffisientene og få ut en verdi på diskriminantscoren. La oss ta et eksempel: Anta at vi har en mann med 20 års utdanning og som er 35 år gammel. Vil denne personen sannsynligvis kjøpe eller ikke kjøpe produktet vårt? Vi regner nå ut diskriminantscoren:

*D=-4,296+0,128*20-0,927*0+0,078*35=0,99*

Vi ser at denne personen får en verdi nær 1, noe som betyr at vi kan predikere vedkommende til å være en sannsynlig kjøper av nyhetskanal. Dersom diskriminantscoren derimot hadde vært nær 0 kunne vi trukket motsatt konklusjon. Den verdien mellom 0 og 1 som skiller de to gruppene er det forhåndsdefinerte kuttpunktet, som for eksempel kan være 0,5. Dersom den aktuelle kundens verdi er over dette kuttpunktet, konkluderer man med at kunden sannsynligvis vil komme til å kjøpe produktet nyhetskanal. Er derimot verdien vi får ut mindre enn det aktuelle kuttpunktet, konkluderer man med at kunden sannsynligvis ikke kommer til å kjøpe produktet.

En slik prediksjonsregel er enkel å håndtere i tilfellet med to grupper. Hadde vi derimot hatt tre grupper hadde dette krevd en mer komplisert utregning.

Tabellen ”structure matrix” viser korrelasjonen mellom hver av variablene som har vært med i analysen og diskriminantfunksjonen. Her ser vi at variabelen inntektskategori er høyere korrelert med diskriminantfunksjonen enn for eksempel variablene kjønn og antall år med utdanning. Dette til tross for at den første av disse tre variablene ikke ble tatt med i den stegvise analysen, mens de to andre ble det. Grunnen til dette er sannsynligvis at variabelen inntektskategori er sterkt korrelert med den første variabelen som ble tatt med i den stegvise analysen, nemlig ”alder i år”.

Structure Matrix

	Function
	1
Alder i år	,822
Inntektskategori(a)	,427
Kjønn	-,386
Antall barn(a)	,320
Antall år med utdanning	,215
Timer fremfor TV en hver dag(a)	-,167
Antall organisasjoner en er medlem av(a)	,074

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.
a This variable not used in the analysis.

Fishers'-koeffisientene som vi finner i tabellen under kan brukes til prediksjon slik som de ustandardiserte koeffisientene. Siden vi i vårt tilfelle bare har to mulige kategorier på den avhengige variabelen, er det likegyldig hvilket sett med koeffisienter vi bruker. Hadde vi hatt flere enn to utfallskategorier, for eksempel ikke kjøp, kanskje kjøp senere og kjøp av nyhetskanal ville Fishers'-koeffisientene vært de mest fornuftige å bruke.

Classification Function Coefficients

	Aksepterer nyhetstilbudet	
	No	Yes
Antall år med utdanning	2,077	2,176
Kjønn	1,985	1,265
Alder i år	,325	,386
(Constant)	-20,851	-24,200

Fisher's linear discriminant functions

8 Faktoranalyse

8.1 Kva er ein faktor?

Faktoranalyse er ein teknikk som vanlegvis blir brukt til datareduksjon. Med datareduksjon meiner vi at vi frå eit stort datasett med mange variablar prøver å gjere datasettet meir handterleg ved å erstatte ei samling av korrelerte variablar med såkalla faktorar. Det kan ofte vere litt vanskeleg å forstå kva ein faktor er for noko, men eit døme kan kanskje gjere ting klarare.

DØME:
For å avgjere kva eigenskapar folk legg vekt på når dei kjøper tannkrem, gjennomfører ein tannkremprodusent ei spørjeundersøking. Folk som blir intervjua får seks påstandar dei skal avgjere i kva grad dei er einige i.

- Tannkremen skal beskytte mot hol.
- Tannkremen må gje kvite tenner.
- Tannkremen må gje sterkare tannkjøt.
- Tannkremen må gje frisk pust.
- Ein viktig fordel med tannkrem er at han hindrar tannrote eller andre typar forfall av tennene.
- Det viktigaste for meg når eg kjøper tannkrem er at tannkremen gjev fine tenner.

Her har vi seks variablar:

X_1: Beskyttelse mot hol.
X_2: Kvite tenner
X_3: Sterkt tannkjøt
X_4: Frisk pust
X_5: Forfall av tanngarden.
X_6: Fine tenner

Dette er dei seks variablane som kan observerast direkte. Men når vi ser nærare på variablane, ser vi at fleire av dei truleg er sterkt korrelerte. Truleg vil personar som synest det er viktig med sterkt tannkjøt, også seie at det er viktig at tannkremen beskyttar mot hol, og at tennene ikkje blir øydelagde på annan måte. Høge korrelasjonar vil vi truleg også sjå mellom variablane X_2, X_4 og X_6.

Når vi har variablar som er korrelerte på denne måten, vil det ofte vere føremålstenleg å identifisere faktorar som reduserer kompleksiteten i datasettet, i staden for å bruke dei observerte variablane direkte.Dersom faktorane blir valde på ein god måte, vil det vere mykje lettare for tannkrem-produsenten å vite kva han skal leggje vekt på i marknadsføringa, enn i tilfellet der han ikkje gjennomfører datareduksjon først. I dette dømet viser det seg, som vi gjekk ut frå, at variablane X_1, X_3 og X_5 er høgt korrelerte, og det same er tilfellet for dei tre andre variablane. I staden for å

bruke X_1, X_3 og X_5 som tre ulike variablar, kan vi innføre ein faktor som vi kallar ”helsemessige omsyn” som ei erstatning for desse tre variablane. På same måten kan vi innføre ein faktor som vi kallar ”estetiske omsyn”, som erstatning for X_2, X_4 og X_6.

Faktorane er altså ikkje noko vi observerer direkte, men eit fellestrekk for fleire variablar som vi analyserer oss fram til vha. faktoranalyse. Når vi brukar faktoranalyse, er det fordi vi ikkje veit kva faktorane er. Målet er å finne ut kor mange faktorar som skjuler seg i datamaterialet.

I dømet ovanfor kunne det verke som vi resonnerte oss fram til kva som var faktorar. Slik vil det vanlegvis ikkje vere. Først etter å ha brukt teknikkar for faktoranalyse, vil det bli tydelegare kva som bør brukast som faktorar. Reknearbeidet omfattar ulike steg, frå metodar for utveljing av antal faktorar, til tolking av faktorane. Eit veldig viktig poeng er likevel at vi må ha færre faktorar enn antal variablar. Elles oppnår vi ikkje det som oftast er poenget med faktoranalyse, nemleg å få til eit datasett som er meir handterleg, og som gjer analysearbeidet lettare.

I dette tilfellet har vi berre seks variablar. Men tenk om vi hadde 60 eller 600 variablar. Då ser vi at det ville vere svært vanskeleg å trekkje ut nødvendig informasjon for å iverksetje tenlege tiltak retta mot potensielle kundar, dersom vi ikkje først gjennomfører datareduksjon og erstattar variablar med faktorar.

8.1.1 Faktoranalyse i marknadsføring

I marknadsføringsfaget blir faktoranalyse gjerne nytta på desse områda:

1. Som metode for å segmentere marknaden i ulike grupper.
2. For å finne ut kva som påverkar etterspurnaden etter eit produkt.
3. Som grunnlag for vidare analyse, t.d. regresjonsanalyse eller clusteranalyse.

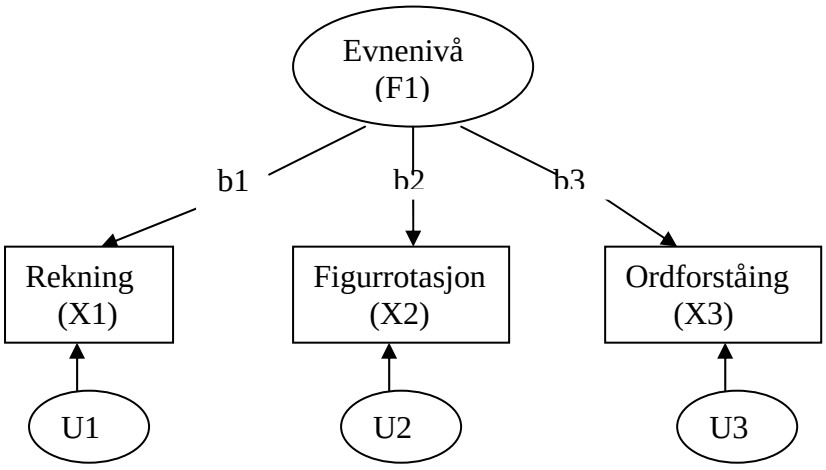
Faktoranalyse innan marknadsføring tek ofte utgangspunkt i ei spørjeundersøkjing. Det er difor viktig at ein tenkjer godt gjennom kva ein spør om, og ikkje minst kva ein vil finne ut. Det er t.d. viktig at spørsmåla blir stilte slik at ein kan skilje ut variablar som er intervallnivå eller på forholdstalnivå, dvs. at variablane blir målte på ein skala der avstanden mellom punkta gjev meining, og at variablane blir tilordna numeriske verdiar. Ein bør difor i teorien unngå å bruke dikotome variablar (ja/nei) eller variablar som er målte på ordinalnivå (dvs. der variaberverdiane er delte inn i ulike kategoriar som er ordna i forhold til kvarandre. T.d. Svært god, God, Dårleg). I praksis ser vi likevel ofte at det blir brukt variablar med ordinalt målenivå i faktoranalyse. Og dette blir heller ikkje sett på som eit problem i tilfelle der variabelen skal reflektere ein faktor som er kontinuerleg.
Innanfor fagfeltet faktoranalyse har vi to hovudteknikkar: Prinsipal komponent-analyse og vanleg faktoranalyse. Desse metodane skil seg ikkje

så mykje frå kvarandre. Prinsipal komponentanalyse blir gjerne brukt når føremålet er å finne fram til det minste antalet faktorar som kan brukast, og der føremålet er å leggje grunnlag for vidare statistisk analyse. Vanleg faktoranalyse (som igjen finst i tre undervariantar) blir meir brukt i tilfelle der ein legg meir vekt på å kome fram til ei god tolking av faktorane. Vanlegvis vil resultata bli såpass like at det ikkje er avgjerande kva for metode ein brukar.

8.2 Faktorteori

Vi har alt teke for oss skilnaden mellom faktorar og variablar. Vi skal bruke eit nytt døme som grunnlag for å skrive litt meir om teorien bak faktoranalyse. Dette dømet er ikkje henta frå marknadsføring.

Gå ut frå at vi ynskjer å seie noko om evnenivået til ein person når det gjeld å løyse teoretiske oppgåver i ein test. Vi tenkjer oss at det er evnenivået som er grunnlaget for korleis ein løyser tre ulike oppgåver i testen. Evnenivået kan vi ikkje måle direkte. Derimot kan vi måle resultatet på dei tre oppgåvene direkte. Dei tre oppgåvene omhandlar rekning, figurrotasjon og ordforståing. Desse tre emna kan vi bruke som variablar, og evnenivået som ein faktor.



Figur 1

Figuren prøver å illustrere at måleresultata for dei tre observerte variablane kan uttrykkjast på følgjande måte:

$$X_1 = b_1F_1 + U_1 \qquad \text{Likning 1}$$

$$X_2 = b_2F_1 + U_2 \qquad \text{Likning 2}$$

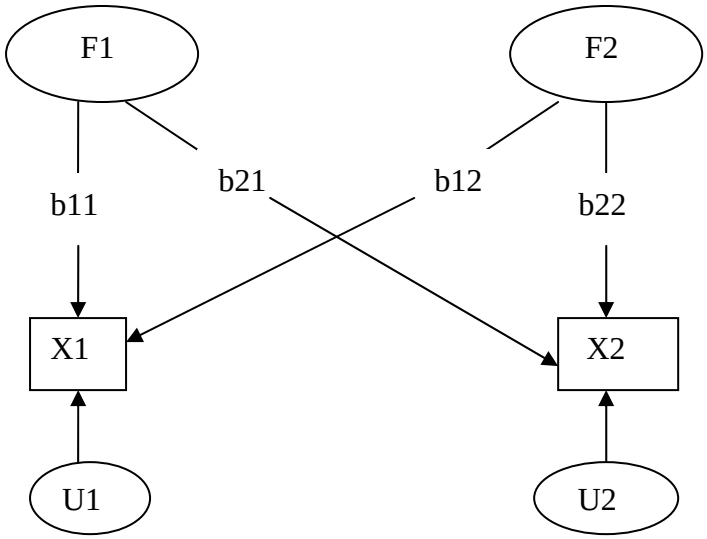
$$X_3 = b_3F_1 + U_3 \qquad \text{Likning 3}$$

Likningane uttrykkjer at nivået på dei tre variablane i figuren, X_1, X_2, og X_3, er ein funksjon av ein felles faktor, F_1, og ein faktor U, som er unik for kvar

av variablane. b er faktorladninga, som viser kor stor betydning F har for variasjonen i nivåa på dei tre variablane. Dersom b_1 er større enn b_2, har faktoren større betydning for *variasjonen* i X_1 enn i X_2. F_1 står for scoren på faktoren (evnenivå). Dette observerer vi ikkje, men vi må tenkje oss at kvar person har ein viss score på faktoren som avgjer nivået til kvar av variablane.

Når vi snakkar om faktorladning og betydning, meiner vi betydning for variasjonen. T.d. kan det vere heilt nødvendig å sjå, lese og skrive for å svare på ein test. Men dersom alle ser godt nok, og er flinke nok til å lese og skrive, vil ikkje dette påverke variasjonen, dvs. dei individuelle skilnadene i resultata. Difor vil ikkje syn og det å kunne lese og skrive kome ut som faktorar i analysen, sjølv om dei er nødvendige føresetnader for resultata.

Måleresultata i ein variabel kan også vere ein funksjon av fleire underliggjande, latente faktorar. I figur 3 har vi eit døme på ein slik modell.



Figur 2

b_{11} står for betydninga av faktor 1 for variasjonen i X_1, b_{21} er betydninga av faktor 1 for variasjonen i X_2. Tilsvarande for b_{12} og b_{22}. Vi får difor desse likningane:

$$X_1 = b_{11}F_1 + b_{12}F_2 + U_1 \qquad \text{Likning 4}$$

$$X_2 = b_{21}F_1 + b_{22}F_2 + U_2 \qquad \text{Likning 5}$$

Det kan visast at korrelasjonen mellom dei to variablane kan skrivast som

$$r_{12} = b_{11}b_{21} + b_{12}b_{22} \qquad \text{Likning 6}$$

Denne likninga gjeld dersom faktorane ikkje er innbyrdes korrelerte.

I praksis veit vi verken talet på faktorar eller faktorladningane på førehand. Desse kan vi finne gjennom å studere korrelasjonane mellom dei observerte variablane. I faktoranalyse går vi ut frå at det er fellesfaktorane (F-faktorane) som bestemmer korrelasjonen mellom variablane, dvs. det er fellesfaktoarane som er årsak til korrelasjonen. Til større betydning ein faktor har for to variablar, til høgare blir korrelasjonen mellom dei. Faktorane som er unike (U-faktorane) for kvar av variablane, påverkar ikkje korrelasjonen mellom variablane.

Faktorladningane (b-ane) er også ukjende på førehand. Desse finn vi fram til ut frå korrelasjonskoeffisientane, som vi reknar ut frå dei svara folk har gitt på spørsmåla dei blir presenterte for. Faktoranalyse kan vi seie går ut på å finne faktorladningane ut frå korrelasjonskoeffisientane, dvs. finne det som står på høgre side i likning 5 ut frå det som står på venstre side. Faktorladningane er eigentleg korrelasjonskoeffisientar mellom faktorane og variablane. Faktorladningane kan difor også vere negative.

DØME: Korrelasjonsmatrise og faktormatrise

Korrelasjonsmatrise

Variablar

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
X1		0,64	0,72	0,48	0	0	0
X2	0,64		0,72	0,48	0	0	0
X3	0,72	0,72		0,54	0	0	0
X4	0,48	0,48	0,54		0,49	0,42	0,56
X5	0	0	0	0,49		0,42	0,56
X6	0	0	0	0,42	0,42		0,48
X7	0	0	0	0,56	0,56	0,48	

Faktormatrise

Faktorar

	F1	**F2**	**h^2**
X1	0,8	0	0,64
X2	0,8	0	0,64
X3	0,9	0	0,81
X4	0,6	0,7	0,85
X5	0	0,7	0,49
X6	0	0,6	0,36
X7	0	0,8	0,64

Eigenvalue:	2,45	1,98
% forklart varians	35	28,3
% total forklart varians:	63,3	

I dette dømet har vi sju variablar. Korrelasjonane mellom variablane står i matrisa til venstre.
Som nemnt er det fellesfaktorane som ligg til grunn for korrelasjonen mellom variablane. I dømet ovanfor er korrelasjonsmatrisa så tydeleg at vi utan vidare ser at dei tre første variablane har ein felles faktor, som dei ikkje deler med dei tre siste variablane. Dette er fordi variablane X_1-X_3 ikkje er korrelerte med variablane X_5-X_7. Tilsvarande har dei tre siste variablane ein felles faktor. Variabelen X_4 ser ut til å ha noko av begge faktorane, sidan variabelen er korrelert med alle dei andre variablane. I praksis vil vi aldri

finne ei så enkel korrelasjonsmatrise, og det kan vere vanskelegare å finne fram til rett antal faktorar.

Resultatet av ei faktoranalyse er ei faktormatrise, som vist til høgre i tabellen. I denne matrisa finn vi faktorane som kolonner, og variablane som rader. Matrisa viser faktorladningane, dvs. kor stor betydning kvar faktor har for variasjonen i nivåa på kvar variabel. Her ser vi at det er to faktorar, dei fire første variablane har ladningar på den første faktoren, og dei fire siste på den andre faktoren. Kolonna h^2 viser kommunalitetane (sjå forklaring neste side) for dei ulike variablane.

Om vi set inn faktorladningane i likning 5 og reknar ut korrelasjonen mellom X_1 og X_4, får vi:

$r_{14} = 0.80*0,60 + 0*0,70 = 0,48$

Dette stemmer med talet i korrelasjonsmatrisa. Det gjer det også for dei andre korrelasjonane. Ofte vil vi ikkje klare å reprodusere korrelasjons-matrisa på denne måten berre med nokre få faktorar. I slike tilfelle legg vi til fleire faktorar. Men sjølv om faktorladningane er valde på ein måte som i størst mogleg grad skal gjere reproduksjon mogleg, er det likevel ikkje sikkert vi klarer å reprodusere korrelasjonsmatrisa perfekt utan å ta med like mange faktorar som variablar. Vi har difor visse kriterie som fortel oss kva som må til for at vi skal ta med ein faktor i modellen.

8.2.1 Kommunalitet, eigenvalue og forklart varians

Dette er tre svært viktige omgrep i faktoranalysen:

Kommunalitet (h^2):
Viser kor stor andel av variasjonen i kvar variabel som fellesfaktorane til saman forklarer. For X_4 ser vi av faktormatrisa på forrige side at kommunaliteten er 0,85, dvs. at F_1 og F_2 forklarer 85 % av variasjonen i variabelen. Resten, dvs. 15 %, er forklart av den unike faktoren U.

Kommunaliteten = summen av kvadrerte faktorladningar. For X_4 har vi:
$0,85 = 0,6^2 + 0,7^2$

Eigenvalue:
Dette er eit mål på kor mykje av variansen i variablane som den enkelte faktoren forklarer. Ein skulle kanskje tru at det var meir naturleg å bruke namnet eigenverdi, men på grunn av at dette omgrepet har ei anna tyding i matematikken, er det vanleg å bruke den engelske nemninga eigenvalue. Vi reknar ut eigenvalue for ein faktor ved å kvadrere faktorladningane i kvar kolonne for seg, og deretter summere desse.
For faktor 1 får vi: $0,8^2 + 0,8^2 + 0,9^2 + 0,6^2 = 2,45$

Forklart varians:

Dette målet heng nøye saman med eigenvalue. Forklart varians finn vi ved å dele eigenvlaue på antal variablar, og rekne om til %.
For faktor 1 får vi: (2,45 / 7) * 100 % = 35 %

Total forklart varians:
Dette = summen av forklart varians for kvar av faktorane.

8.3 Dei tre hovudstega i faktoranalyse

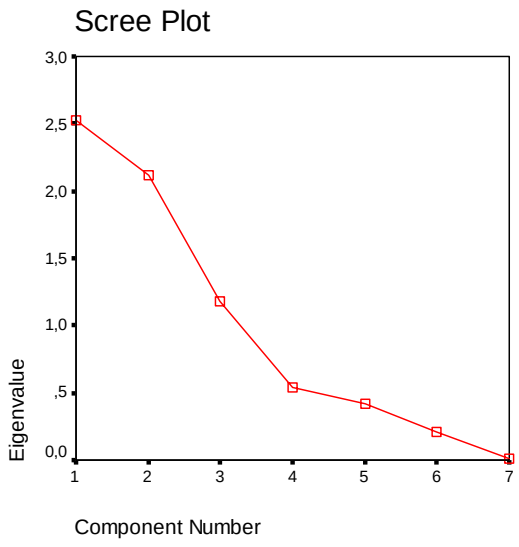
8.3.1 Utrekning av faktorar

Vi skil mellom eksplorerande og bekreftande faktoranalyse. I bekreftande faktoranalyse vel vi talet på faktorar på førehand, og testar hypoteser ved hjelp av teknikkar for faktoranalyse. For å gjennomføre bekreftande faktoranalyse treng vi ein anna dataverkty enn SPSS. I eksplorerande faktoranalyse, som vi skal sjå på her, veit vi ikkje talet på faktorar på førehand, men brukar bestemte metodar for å fastsetje kor mange faktorar vi skal ta med.

Den mest brukte metoden for uttrekning av faktorar, er å velje ut dei faktorane som har eigenvalue > 1. Dette blir ofte omtalt som Kaisers kriterium.

Eit alternativ er å sjå på ei grafisk framstilling, kalla Scree plot, og velje faktorar skjønnsmessig ut frå dette plottet. Plottet viser faktorane på førsteaksen, og Eigenvalue på andreaksen. Figuren nedanfor viser døme på eit scree-plot. Vi ser at grafen har eit markert brot etter at vi har teke med fire faktorar, dvs. det er lite å tene på å ta med fleire faktorar. Her vel vi difor fire faktorar.

Det mest vanlege er likevel å bruke Kaisers kriterium for å velje ut faktorar.



8.3.2 Rotering

Målet med rotering er å gjere faktorane meir forståelege. Faktorane har ofte ladningar med varierande styrke på alle variablane, og faktormatrisa blir difor vanskeleg å tolke. Ved å bruke rotasjon, prøver vi å få til ein situasjon der faktorane har høge ladningar på nokre få variablar, og ladningar nær null på dei andre variablane.

Den roterte løysinga forklarer like stor del av variasjonen som den opphavlege, uroterte løysinga. Men forklaringskrafta til kvar einskild variabel har gjerne endra seg.

8.3.3 Tolking av faktorane

Etter å ha trekt ut faktorar, og gjennomført rotasjon, må vi prøve å tolke faktorane. Kva er det faktorane eigentleg seier noko om? Dersom vi har oppnådd det vi ynskjer med roteringa, er det vanlegvis nokså greitt å gje ei tolking av faktorane.

9 Faktoranalyse i SPSS

Hent først fram fila **hatco.sav.** Fila viser ei marknadsundersøking som er gjort av firmaet Hatco. På grunnlag av 14 variablar er målet å trekkje ut faktorar som er viktige for kundane si oppfaning av firmaet.

9.1 Er krava til å gjennomføre faktoranalyse oppfylte?

Før ein gjennomfører faktoranalyse på datasettet, bør ein finne ut om datasettet verkeleg eignar seg til å analysere ved hjelp av faktoranalyse. Dersom t.d. variablane i datasettet er tilnærma ukorrelerte, vil vi måtte bruke like mange faktorar som variablar, og det er ikkje poeng i å bruke faktoranalyse. På den andre sida kan vi også ha så høge korrelasjonar mellom nokre av variablane at vi får problem med multikollinearitet. I slike tilfelle må vi anten kombinere desse variablane på ein måte slik at vi unngår multikollinearitet, eller fjerne ein eller fleire av variablane før vi gjennomfører heile analysen. Sjølv om vi må kutte ut nokre av variablane, kan likevel faktoranalyse vere eit svært nyttig hjelpemiddel.

Vi har to testar for å finne ut om faktoranalyse bør brukast. Det er KMO-testen og Bartlett`s test of sphericity. Testane finn de under

Analyze
Data reduction
Factor
Descriptives.
Markér for KMO and Bartlett`s test of sphericity.

Både KMO og Bratlett`s test of Spericity seier noko om kor godt eigna datasettet er for å gjennomføre faktoranalyse. Bartlett-testen testar ein nullhypotese om at korrelasjonsmatrisa er ei identitetsmatrise (dvs. ei matrise der korrelasjonane mellom alle variablane = 0). I eit slikt tilfelle vil ein ikkje kunne gjennomføre faktoranalyse. Målet må difor vere å forkaste H_0.

Det blir rekna ut ein KMO-verdi både for kvar variabel, og for alle variablane samla (dvs. som summen av dei individuelle verdiane). Den samla verdien bør vere over 0,5 for å bruke faktoranalyse. Dersom dette kravet ikkje er oppfylt, må vi sjå på KMO utrekna for kvar enkelt variabel, fjerne variabelen med lågast KMO-verdi, og gjennomføre faktoranalysen på nytt. Denne prosedyren bør vi gjenta til KMO er > 0,5.

KMO-verdiane for kvar enkelt variabel finn vi i den såkalla anti-image-matrisa, som vi finn under

Analyze
Data reduction
Factor
Descriptives.
Markér for Anti-image.

DØME:
Vi ynskjer å finne ut om vi kan ”redusere” datasettet ved å bruke faktoranalyse på variablane X1-X7. Vi kunne sjølvsagt gjort ein tilsvarande test på alle 14 variablane, men for å unngå altfor store og uhandterlege matriser i dette notatet, har vi plukka ut berre nokre av variablane.

Frå menyen vel vi

Analyze
Data reduction
Factor

I dialogboksen markerer vi for variablane X1-X7, og flytter desse over i boksen Variables. Under alternativet Descriptives markerer vi for KMO and Bartlett`s test of sphericity og Anti-image. Dette gjev m.a. desse resultata:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,446
Bartlett's Test of Sphericity	Approx. Chi-Square	567,541
	Df	21
	Sig.	,000

Vi ser at KMO-estimatoren er < 0,5, og faktoranalyse er difor ikkje så godt eigna til å gjennomføre på variablane X1-X7. Dette gjeld sjølv om Bartlett-testen er signifikant. Men kanskje kan vi bruke faktoranalyse om vi fjernar ein av dei sju variablane, og og gjennomfører analysen på dei seks gjenverande variablane? Regelen er at ein skal fjerne variabelen med lågast KMO-verdi. Desse verdiane er markerte med (a) i anti-image-matrisa nedanfor. I dette tilfellet er det variabelen X5 (Service) som blir fjerna. Dette gjer vi ved å flytte denne variabelen frå boksen Variables til dialogboksen.

Anti-image Matrices

		Delivery Speed	Price Level	Price Flexibilty	Manufacturer Image	Service	Saleforce Image	Product Quality
Anti-image Covariance	Delivery Speed	,028	,028	,002	,015	-,025	-,006	-,002
	Price Level	,028	,032	,022	,014	-,026	-,005	-,020
	Price Flexibility	,002	,022	,608	,044	-,011	-,040	,086
	Manufacturer Image	,015	,014	,044	,347	-,015	-,275	-,018
	Service	-,025	-,026	-,011	-,015	,023	,005	,010
	Salesforce Image	-,006	-,005	-,040	-,275	,005	,371	-,044
	Product Quality	-,002	-,020	,086	-,018	,010	-,044	,623
Anti-image Correlation	Delivery Speed	,344(a)	,957	,018	,149	-,978	-,060	-,016
	Price Level	,957	,330(a)	,155	,134	-,975	-,045	-,141
	Price Flexibility	,018	,155	,913(a)	,095	-,091	-,085	,140
	Manufacturer Image	,149	,134	,095	,558(a)	-,173	-,766	-,039
	Service	-,978	-,975	-,091	-,173	,288(a)	,052	,088
	Salesforce Image	-,060	-,045	-,085	-,766	,052	,552(a)	-,092
	Product Quality	-,016	-,141	,140	-,039	,088	-,092	,927(a)

a Measures of Sampling Adequacy(MSA)

Vi gjennomfører same prosedyre som ovanfor med dei seks gjenverande variablane. Dette gjev følgjande verdiar for KMO og Bartlett`s Test:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,665
Bartlett's Test of Sphericity	Approx. Chi-Square	205,965
	df	15
	Sig.	,000

No ser vi at krava til å gå vidare med faktoranalysen er oppfylte. KMO er > 0,5, og signifikans-sannsynet for Bartlett-testen er tilnærma lik 0. Vi tek ikkje med anti-image-matrisa her, fordi denne blir brukt til å fjerne variablar som gjer det lite føremålstenleg å bruke faktoranalyse. Her har vi ingen slike variablar. Vi kan difor gå vidare til sjølve analysen.

9.2 Gjennomføring av faktoranalyse

Det varierer frå oppgåve til oppgåve kor detaljert informasjon vi ynskjer. Vanlegvis vil vi ta med matriser for korrelasjon mellom variablane, oversikt over kommunalitetane, eigenvalues, forklart varians, faktormatriser og roterte faktormatriser.

Korrelasjonsmatrise

Det er korrelasjonar mellom variablane som legg grunnlag for faktoranalyse. Dersom datasettet ikkje inneheld sett av variablar som er korrelerte med kvarandre, vil vi heller ikkje kunne bruke faktoranalyse. Korrelajonsmatrisa treng vi ikkje direkte i analysen, men ofte blir det stilt spørsmål som krev at matrisa er tilgjengeleg.

Matrisa finn vi ved å gå inn i menyane

Analyze
Data reduction
Factor
Descriptives.
Markér for Coefficients og Significance levels. Dette gjev denne matrisa:

Correlation Matrix(a)

		Delivery Speed	Price Level	Price Flexibility	Manufacturer Image	Salesforce Image	Product Quality
Correlation	Delivery Speed	1,000	-,349	,509	,050	,077	-,483
	Price Level	-,349	1,000	-,487	,272	,186	,470
	Price Flexibility	,509	-,487	1,000	-,116	-,034	-,448
	Manufacturer Image	,050	,272	-,116	1,000	,788	,200
	Salesforce Image	,077	,186	-,034	,788	1,000	,177
	Product Quality	-,483	,470	-,448	,200	,177	1,000
Sig. (1-tailed)	Delivery Speed		,000	,000	,309	,223	,000
	Price Level	,000		,000	,003	,032	,000
	Price Flexibility	,000	,000		,125	,367	,000
	Manufacturer Image	,309	,003	,125		,000	,023
	Salesforce Image	,223	,032	,367	,000		,039
	Product Quality	,000	,000	,000	,023	,039	

Første del av matrisa viser dei bivariate korrelasjonane mellom ulike variablar. Siste del av matrisa viser om korrelasjonane er signifikante.

9.2.1 Kommunalitetar, eigenvalues og forklart varians

Informasjon om desse storleikane vil vi få fram automatisk når vi gjennomfører faktor-analysen. Vi slepp altså å markere for ekstra informasjon i undermenyane.

Kommunalitetar

Omgrepet er forklart tidlegare i kompendiet.

Communalities

	Initial	Extraction
Delivery Speed	1,000	,658
Price Level	1,000	,580
Price Flexibility	1,000	,646
Manufacturer Image	1,000	,882
Salesforce Image	1,000	,872
Product Quality	1,000	,616

Extraction Method: Principal Component Analysis.

Under tabellen ser vi at det står Extraction Method: Principal Component Analysis. Det er altså prinsipal komponentanalyse (PCA) som blir brukt i SPSS dersom vi ikkje markerer for at vi vil bruke ein annan uttrekningsmetode. I PCA brukar vi namnet komponent i staden for faktor. Dersom vi ynskjer å bruke ein annan metode enn PCA, kan vi gjere det på denne måten:

Analyze
Data reduction
Factor
Extraction

Forklaring til kolonnene i tabellen:

Initial:

Dette er estimat for varians i kvar variabel som er forklart av alle faktorar, både fellesfaktorar og unike faktorar. I PCA blir alltid desse estimata lik 1. Det er ikkje tilfellet for vanleg faktoranalyse, der ein utelet forklaringskrafta frå faktorane som er unike for kvar variabel. Kommunaliteten vil såleis bli < 1 også i kolonna Initial.

Extraction:

Estimat for variansen i kvar variabel som er forklart ut frå faktorane i faktorløysinga, dvs. etter at vi har gjennomført datareduksjon og berre står att med faktorar som har eigenvalue > 1 (sjå neste avsnitt).

Eigenvalues og forklart varians
Informasjonen om storleikane blir presenterte på denne måten:

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,513	41,892	41,892	2,513	41,892	41,892	2,370	39,497	39,497
2	1,740	28,992	70,883	1,740	28,992	70,883	1,883	31,386	70,883
3	,597	9,958	80,842						
4	,530	8,826	89,668						
5	,416	6,929	96,596						
6	,204	3,404	100,000						

Extraction Method: Principal Component Analysis.

Initial Eigenvalues:
Eigenvalue viser den totale variansen forklart av kvar faktor. Den totale mengda som skal forklarast set vi lik antal komponentar (faktorar), som i utgangspunktet er lik antal variablar. I dette tilfellet skal vi altså forklare ein varians på 6. Av denne summen kan vi forklare 2,513 med den første faktoren. Dette = 41,892 % av den totale mengda (6).
Kolonna for cumulative viser den samla forklaringskrafta for dei faktorane vi har inkludert. Etter å ha inkludert to faktorar har vi forklart 41,892 % + 28,992 % = 70,883 % av den totale variansen.

Som nemnt tidlegare er det vanleg å ta med alle faktorane som har eigenvalue > 1, dvs dei faktorane som forklarer meir av variansen enn ein standardisert variabel.

Extraction sums of squared Loadings:
Denne kolonna viser kor mykje av variansen dei faktorane som vi tek med i den endelege løysinga, forklarer. I dette tilfellet endar vi opp med to faktorar dersom vi brukar kriteriet om eigenvalue > 1. Desse tre forklarer til saman 70,883 % av variansen. For å få fram denne kolonna av tabellen, må vi bruke kommandoen

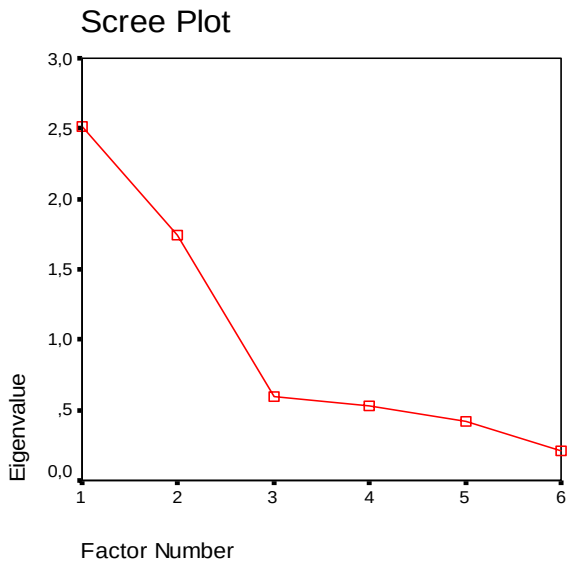
Analyze
Data reduction
Factor
Extraction
Unrotated factor solution.

Rotation sums of squared loadings:
Føremålet med å rotere er å gje faktorane ei meir intuitiv tolking. Den forklarte variansen vil samla sett ikkje endre seg, men vi vil sjå at fordelinga av forklarings-krafta vil bli jamnare etter rotering, og det vil vanlegvis bli

lettare å tolke faktorane. For å få med denne delen av tabellen ovanfor må vi bruke kommandoen

Analyze
Data reduction
Factor
Rotation
Markér for Varimax under Method og Rotated Solution i Display-boksen.

Scree-plot
Dette er ein illustrasjon av eigenvalues til dei ulike faktorane. Plottet kan som nemnt tidlegare brukast som eit hjelpemiddel i uttrekninga av faktorar. I dette dømet ville vi velje ut tre faktorar, medan vi med eigenvalue-kriteriet fann at vi skulle velje to. Vi held oss til eigenvalue i tilfelle der dei to kriteria skil seg frå kvarandre.



Scree-plottet blir vist ved å bruke menyane

Analyze
Data reduction
Factor
Extraction
Markér for Scree plot i Display-boksen.

Faktormatrise (komponentmatrise)
Denne matrisa viser faktorladningane til dei ulike faktorane (komponentane). Faktor-ladningane er korrelasjonar mellom faktorane og

variablane. Denne matrisa vil bli vist i SPSS når vi gjennomfører faktoranalyse berre med kommandoen
Analyze
Data reduction
Factor
Vel ut variablane vi ynskjer å bruke, og trykkjer OK.

Component Matrix(a)

	Component	
	1	2
Delivery Speed	-,627	,514
Price Level	,759	-,068
Price Flexibility	-,730	,337
Manufacturer Image	,494	,798
Salesforce Image	,425	,832
Product Quality	,767	-,168

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Rotert faktormatrise (komponentmatrise)

Ved å markere for Varimax og Rotated solution i Rotation-menyen får vi denne matrisa:

Rotated Component Matrix(a)

	Component	
	1	2
Delivery Speed	-,787	,194
Price Level	,714	,266
Price Flexibility	-,804	-,011
Manufacturer Image	,102	,933
Salesforce Image	,025	,934
Product Quality	,764	,179

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 3 iterations.

Denne matrisa viser faktorladningane etter at vi har rotert matrisa. Målet med rotasjonen er at vi skal få fleire høge ladningar, og fleire ladningar som er nær null. På denne måten blir faktorane vanlegvis enklare å tolke.

I dette dømet ser vi at variablane leveringstid, prisnivå, prisfleksibilitet og produktkvalitet har høge ladningar på faktor 1. Det er litt vanskeleg å setje eit godt namn på ein faktor som omfattar så mykje, men vi ser at alle variablane i alle fall har litt med oppfatning av karakteristika ved selskapet Hatco å gjere.

Faktor to har høge ladningar på dei to gjenverande variablane. Vi kan kalle faktoren for image.

Litteraturliste

Fugleberg O. og Kristianslund I (1995) Innføring i regresjonsanalyse og multivariate metoder. Bedriftsøkonomens Forlag.

Gujarati D. N. (2003) Basic Econometrics. Mc Graw Hill.

Malhotra & Birks (2003), Marketing Reasearch. An Applied Approach. Pearson Education.

Mogensen B. (2001) SPSS - Basismodul. Kursdokumentasjon. Capture Data AS.

Selnes (1999). Markedsundersøkelser. Tano Aschehoug.

SPSS (2000). Market segmentation Using SPSS. Kurshefte.

Ulleberg og Nordvik (2000) *Faktoranalyse*. Tapir Forlag 2000

Wenstøp F. (1997) Statistikk og dataanlyse. Universitetsforlaget.

Wonnacott T. H. og Wonnacott R. J. (1990) Introductory Statistics for Business and Economics. John Wiley & Sons.

Filer:

”tabell avling og gjodsel.xls”	Kilde: Wonnacott T. H. og Wonnacott R. J. (1990). Tabell 11-1. Fiktive data.
	Innhold: Styrt eksperiment. Avlingsmengde (bu/acre) og gjødselsmengde (lb/acre).
”Demosav.sav”	Kilde: SPSS. Opprinnelse ukjent. Mulig fiktive data.
	Innhold: Forbruksdata på individnivå.
”Banknor.sav”	Kilde: Capture Data AS. Opprinnelse ukjent. Mulig fiktive data.
	Innhold: Et utvalg fra amerikansk bankvesen. For hver ansatt er det oppgitt årslønn i dollar ved ansettelse og ved utfylling av spørreundersøkelsen. Andre variabler er kjønn, hudfarge, utdannelsestid, alder, yrkeskategori i banken etc.
”Gujarati tab 6 3.sav”	Kilde: Economic Report of the President, 1999, Table B-17, p. 347.
	Innhold: Aggregert totalt privat forbruk i USA fordelt på kategorier. Alt målt i milliarder 1992-dollar.
”Hatco.sav”	Kilde: Ukjent. Innhold:
”Nyhetskanal.sav”	Kilde: SPSS. Opprinnelse ukjent.
	Innhold: En undersøkelse der respondentene svarer på om de kan tenke seg å motta nyheter interaktivt (via kabel) til sitt eget TV-apparat.