Kvantitativ metode

# 1 Introduksjon: grunnleggende konsepter, og dataanalyse i R

Variabler

* Kvalitative metoder er basert på antagelsen om at egenskaper kan kodes inn i variabler.
* For å være interessante må variabler variere, og hensikten med kvantitative metoder er å redegjøre for hva variasjonen skyldes.
* Vi skiller ofte mellom
  + Avhengig variabel
  + Uavhengig variabel
* Forskjellige variabler kan ta uklike typer verdier:
  + Kontinuerlige
  + Kategorisk/faktor
* Et datasett er en samling av en eller flere observerte skårer på en eller flere variabler.

Utvalg og populasjoner

* Populasjon: de du ønsker å uttale deg om.
* Utvalg: et relativt lite antall observasjoner trukket fra populasjonen.
* Statistikk: en egenskap (verdi) regnet ut på bakgrunn av et utvalg.
  + Beskrivende statistikk: kvantifiserer egenskaper ved utvalget
  + Slutningsstatistikk: metoder som lar oss trekke slutninger om populasjonen basert på data

Observasjonelle vs. eksperimentelle studier

* Eksperimentell studier
* Observasjonelle studier

Fordelinger og beskrivende statistikk

Grafer av kategoriske variabler

* Frekvensfordeling: en liste mulige verdier en variabel kan ta, og antall observasjoner som tar samme verdi.
* Kategoriske variabler blir typisk avbildet med stolpediagram.

Grafer av kontinuerlig data

* Histogram: del rangen inn i intervaller, tell antall observasjoner som faller i hvert intervall, og lag et stolpediagram.

Hvordan beskrive fordelinger

* Modalitet
* Skjevhet
* Kurtosis

Sigma notasjon

* For å vise til summen av et sett av elementer bruker vi det greske sigma (sum) tegnet:
  + Øvre og nedre grense kan være definert, og da er det mer presist

Mål på sentralitet

* Mode
* Median
* Gjennomsnitt

Mål på variabilitet

* Varians: mål på variabilitet rundt den forventede verdien (gjennomsnittet på en fordeling)
* Range: maksimum verdi – minimum verdi

Prosentiler

* Prosentil: den p’te prosentil er den verdien som er gitt prosent er mindre eller lik.
* Kvartil: verdien som deler de sorterte datasettet inn i fire like store grupper.
* Interkvartil range (IQR): 75. prosentil – 25. prosentil (Q3-Q1)

Standardavvik

* Egenskaper til standardavviket
  + Et spredningsmål med samme enheter som variabelen
  + Kan tolkes som «gjennomsnittlig avvik fra gjennomsnittet»
* Varians og standardavvik er lite robust

Uteliggere

* Uteligger: en observasjon som ligger langt i verdi fra de øvrige observasjonene

Talls oppsummering av en fordeling

* Boxplot: er ofte en god måte å vise viktigere egenskaper ved fordelingen.

Sannsynlighet

* Et stokastisk utfall er tilfeldig, men følger en regularitet over mange repetisjoner.
* Sannsynlighet kan defineres som andelen ganger et gitt utfall forekommer over et stort antall repetisjoner.
  + Følgelig tar sannsynlighet er verdi mellom 0 og 1.

Sannsynlighetsfordelinger for kontinuerlige utfall

* En sannsynlihetstetthet er en funksjon som viser den relative forekomsten av utfall av en stokastisk variabel som har kontinuerlige verdier.
  + Det totale arealet under kurven er lik 1.
  + Ingen verdier er negative.
  + ---

Normalfordelingen

* Normalfordelingen er kanskje den viktigste fordelingen innen statistikk.

Statistikker vs parametere

* Vi antar ofte at egneskape vi er interessert i følger en bestemt fordeling i populasjonen, og fra denne fordelingen trekker vi et utvalg.
* Verdiene som bestemmer formen på fordelingen i populasjonen kalles parametere
  + Parameter: sanne vurdering av egenskap.
* Verdier er beregnet i et konkret utvalg kalles statistikker.
* Deskriptive statistikker vil ta ulikeverder i forskjellige utvalg trukket fra den samme populasjonen

Arealer av normalfordelingen

* Alle normalfordelinger, uansett form har egenskap:
  + Ca 68% av fordelingen ligger innenfor ett standardavvik fra snittet.
  + Etc.
* Standardavvik fra snittet

z-skåre

* Z-skåre: omregning av variabelen til enheter som utgjør antall standardavvik fra gjennomsnittet.

Introduksjon til R/Rstudio

* Denne uken skal vi i R se på:
  + Introduksjon til visualisering med ggplot2
  + Grunnleggende datahåndtering i R med dplyr
  + Utregning av beskrivende statistikker

Tibbles og data frames

* Tibbles er en moderne versjon av data-frames. Tenk på dem som en variabel som lagrer et excel-ark.
* Dataark: tibble

Pipe operator

* Pipe operatoren, skrevet som %>%, er en viktig operator i magrittr-pakken. Den tar resultatet fra en funksjon og sender den …
  + Tar noe, sender det videre

Dplyr pakken

* Dplyr er en R-pakke som implementerer en grammatikk…
* filter()
* select()
* mutate()
* arrange()
* summarise()

# 2 Bivariat regresjon

Assosiasjonen mellom variabler

Spredningsplot

* spredningsplot: en figur der hvert par med verdier på to variabler plottes som et punkt i planet.
  + Uavhengig variabel utgjør typisk x
  + Avhengig variabel utgjør typisk y-verdien.

Kovarians

* Kovarians: er en statistikk der grad av samvariasjon mellom to variabler kvantifiserer.
* Summen fra punkt 1 til n
* cov(x,y) ,
* bidraget til kovariansen er stor om punktet ligger langt vekk fra både gjennomsnitt av x og y i 2. kvadrant.
* Kovariansen er avhengig av skalaen, og derfor pleier vi å regne om kovariansen til korrelasjon

Pearson correlation

* Pearson-korrelasjon:
  + Utgjør en standardisering av kovariansen.
  + Er definert som kovariansen delt på den største verdien denne kovariansen kan ta
  + Er den vanligste

Forklart varians

* Forklart varians (r^2): andelen varians to variabler har til felles
  + Et tall mellom 0 og 1 (siden korrelasjonen er mellom -1 og 1).
  + Kan si noe om hvor mye korrelasjon den ene variabelen kan forklare av variansen
* Korrelasjon er derfor et mål på effektstørrelse

Eksempel: RWA og åpenhet til erfaringer

* RWA skalaen var utviklet av Altemeyer i 1981
* Folk med høye skårer på RWA underkaster seg i høy grad autoriteter de oppfatter som etablerte og legitime. De holder seg til samfunnskonvensjoner og normer

Korrelasjoner er ikke robuste

* Uteliggere kan ha en veldig stor effekt på korrelasjoner.
* Korrelasjoner er påvirket av range
  + Typsisk synker korrelasjonen ettersom rangen til variablene miner.
* Bakenforliggende variabler kan påvirke korrelasjonen.
  + Konfunderende variabler
  + suppressorvariabler

Korrelasjon og linearitet

* pearson korrelasjoner kun et mål på grad av lineært forhold mellom variablene.
  + Fanger kun opp en viss type sammenheng.

Bivariat regresjon

* Hva gir en regresjonsmodell oss?

Rette linjer

* En rett linje er unikt definert av to parametere:
  + b0: konstantleddet/skjæringsverdi
    - y-verdien når x=0
  + b1: stigningstallet (slope)
    - hvor mye endrer y-verdien seg når x øker med en enhet.

Lesing og IQ

* tenkt sammenheng
* vår modell av IQ består av en systematisk og lineær sammenheng med lesing, og en stokastisk (tilfeldig) komponent.

Regresjonslinjen

* forventet verdi
* gjennomsnittet av en variabel
* en regresjonsfunksjon beskriver hvordan gjennomsnittet til en avhengig variable gølger verdiene til en uavhengig variabel
* forventet y-verdi gitt den observerte verdien av x
* om man tenker på en person og man får vite at det er en basketspiller vil man endre sin forventede verdi

bivariat regresjon

* vi har en avhengig (y) og en uavhengig (x) variabel.
* Den forventede verdien gitt en x verdi basert på de estimerte regresjonskoeffisientene er gitt ved: \*formel\*
* b0 er den ‘sanne’ verdien i populasjonen, mens den estimerte verdien basert på et utvalg med en ‘hatt’
  + hatt over variabel = estimert
* i uttrykket til den forventede verdien tie en gitt person ikke inngår i , fordi den forventede verdien er 0.

tolkning av ustandardiserte regresjonskoeffisienter

* konstantleddet

bivariat regresjon i R

* bruker ‘summary()’
* lm : linear model

prediksjon

* med de estimerte regresjonskoeffisisentene kan du sette oppp uttrykket for den predikerte skåren for den uavhengige variabelen (predicted equation).
* Predikere verdi på den avhengige for hvert nivå av den uavhengige.
* Residuale: forskjellen: observerte-estimerte

Ekstrapolering

* Prediksjon utenfor rangen av den uavhengige variabelen.
* Husk at modellen bare er gyldig i et begrenset intervall, vi kan f.eks. ikke bruke den til å predikere forventet RWA skåre for noen med openness på 15.

Sentrering av uavhengige variabler

* Sentrering av uavhengige variabler innebærer å trekke verdi fra skårene, ofte gjennomsnittet. Dette endrer skalaen, men påvirker ikke enhetene eller stigningstallet.
  + Sentrering kan bandt annet gjøre b0 mer tolkbar.
* Tar en individuell skåre og trekker fra gruppeskåren
* De som har gjennomsnittlige skårer, har null
* Interceptet her er forventet verdi for de som har gjennomsnittlig åpenhet

Logikken bak standardiserte koeffisienter

* Sette enhetene på en sammenlignbar skala

Standardiserte vs ustandardiserte koeffisienter

* b1:

hvordan få standardiserte koeffisiener fra R

* lm.beta

Estimering av koeffisienter

Identifikasjon

* i underidentifiserte systemer er det ikke nok info til å unikt estimere verdier for modell parametrene.
* I akkurat identifisert systemer er det ett unikt sett med estimater.
* I overidentifiserte systemer (slik vi vanligvis har), kan vi ikke finne en løsning som passer perfekt.

Tilpasse en linje til data

* Residual: differansen mellom den observerte og predikerte verdien.
* Linjen hvor man får minst residualer vil passe best på et vis.

Minste kvadraters estimering

* Det minste kvadraters estimatet for b0 og b1 er verdiene b0 og b1 med hatter som minimerer summen av de kvadrerte residualene.
* SSE
* Summen blir minst mulig
* Det kan vises at minste kvadraters estimater av regresjonskoeffisientene kan regnes ut ved disse uttrykkene (kalles ofte normal equations).

Kvadratsummer (sum of squares)

* TSS= SSM +SSE
* TSS: beste gjetning om man ikke vet noe om x
* SSE: beste linja, men det gråe arealet er *uforklart*
* SSM: *forklart* av den uavhengige variabelen i modellen

Betinget varians er mmidnre enn marginal varians

* Marginal varians: varians rundt gjennomsnittet. Dette er varians vi ikke kan forklare før vi tilpasser regresjonsmodellen. Avvik fra den blå linjen uforklart.
* Betinget varians: variasjon rundt regresjonslinjen ... Avvik fra den grønne linjen uforklart.

Forklart varians

* Kall summen av kvadratene residualer i den naive midellen til venstre E1, og summen av kvadrerte residualer i modellen til høyre E2.
* r^2 er et mål på nedgang…

ANOVA() funksjonen

* du kan skrive ut SSM og SSE ved å bruke anova() funksjonen i R.

Antagelser i regresjon

1. forholdet mellom x og y er lineært.
2. feilvariansen er normalfordelt, og har samme varians for alle nivåer av X.
3. det er ingen ekstreme uteliggere.
4. observasjonene er uavhengige.
5. den uavhengige variabelen er måt uten feil.

# 3 Multippel regresjon

Multippel regresjon (MR)

* partielle regresjonskoeffisienter
* statistisk kontroll
* supressoreffekter
* forklart varians i MR
* partiell korrelasjon
* polynomisk regresjon
* innflytelse

Intro til MR

Hvorfor ha mer enn en uavhengig variabel?

1. mer nøyaktig prediksjon

* isolere effekten av én uavhengig variabel
* den avhengige variabelen er hovedfokus.
* F.eks. predikere bedre hvem som dropper ut av skolen.

1. isolere effekten av en uavhengig variabel (statistisk kontroll)

* interessen er i en eller flere spesifikke uavhengige variabler.
* F.eks. øker fattigdom risikoen fra frafall?

1. forstå den samlede effekten av flere uavhengige variabler

* flere uavhengige variabler kan samlet påvirke utfallet på kompliserte måter.
* f.eks. er effekten av fattigdom på frafall sterkere for barn av enslige foreldre (interaksjon).
* F.eks. fører fattigdom til frafall fordi foreldrene i fattige familier har mindre tid til å følge med på barna (mediering).

Multippel regresjon

* Har vi to uavhengige variabler x1 og x2, refererer vi til person i’s skåre på disse variablene som x1i og x2i
* Parametrene (bj) kalles partielle regresjonskoeffisienter.
  + De tolkes som forventet endring i Y som følger av at Xj endres en enhet, men alle andre avhengige variabler holdes konstant
* Hvorfor inkluderer ikke forventet inntekt epsilon i? fordi vi har en antagelse om hvordan støy er fordelt. Forventninger ligger på linjen, epsilon vil være 0. forventet y-verdi ligger på regresjonslinjen. Den er normalfordelt rundt 0.

Multippel regresjon i R

Mod1 🡨 lm(Inntekt ~ Utdanning+Erfaring)

Summary(mod1)

Tolkning

* Hva er den forventede inntekt for et individ uten utdanning og uten erfaring?
  + Intercept – estimate
* Hvis jeg holder erfaring konstant, hvor mye øker inntektene for hvert år utdanning?
  + Utdanning – estimat

Geometrisk representasjon

* Med to uavhengige variabler utgjør regresjonsmodellen (for den forventede Y-verdien) et plan i tredimensjonalt rom.
* Modellen er et plan

Hvilket regresjonsplan korresponderer til utskriften?

* Til høyre siden trening har negativt stigningstall

Matrix scatterplot

Pairs(cbind(inntekt2, …

R-squared (forklart varians) i multippel regresjon

* Mål på nedgangen i proporsjonen av uforklart varians når vi går fra en modell til en annen.
* Mål på nedgangen eller andelen av uforklart varians når vi går fra en modell til en annen.

Justert R-squared

* Du kan forklare varians ved tilfeldighet, særlig dersom du har mange uavhengige variabler og et lite utvalg.
* Justert r-squared er en statistikk som korrigerer r-squared for utvalgsstørrelse og antall uavhengige variabler i modellen.
  + Når n er stor og p er liten, så er r-squared og justert r-squared praktisk talt identiske.
  + r-squared øker alltid om du legger til flere uavhengige varibaler, men justert r-squared kan synke.
* Estimat på hvor mye den faktisk kan forklare som tar høyde for hvor mange faktorer og prediktorer i modellen

R-squared når prediktorene er ukorrelerte

* Når de uavhengige variablene er ukorrelerte forklarer de «separate proporsjoner» av variansen i den avhengige, og forklart varians i en modell med begge er summen av hva de kan forklare enkeltvis.
* Samlede bidraget av to variabler: Summen av det de kan forklare enkeltvis
  + Men ikke om de er korrelerte!

Økologisk mat

* Er det en helsegevinst av å spise økologisk mat?
  + Folk som spiser mer økologisk mat er kanskje mer opptatt av helse og kosthold. De er systematisk forskjellige fra populasjonen.
* Hvis vi plotter folks inntak av økologisk mat mot deres helse, ser det ut til at høyere nivåer av økologisk mat er forbundet med bedre helse.

Statistisk kontroll

* Problemet med tredjevariabler
* Korrelasjonen reflekterer en kausal effekt av utdanning
* Spuriøs korrelasjon
* Utenforliggende tredjevariabler kan skape en spuriøs korrelasjon
* Spuriøs: falsk, brukt hovedsakelig innen epidemiologi for å beskrive en relasjon som kan virke kausal, men ikke er det.
* Utenforliggende variabel (extraneous): andre variabler enn den uavhengige variabelen i et eksperiment som påvirker …

Konfundering

* Konfunderende variabler korrelerer altså med både eksponeringen og utfallet.
  + Er du interessert i sammenhengen mellom tidlig alkoholdebut og senere alkoholkonsum er sensation seeking er en mulig konfunderende variabel.

Statistisk kontroll

* I eksperimentelle studier fjerner vi effekten av konfunderende variabler ved å holde dem konstante.
* I observasjonelle studier trenger vi også en måte å holde en «konfunderende variabel konstant»
  + Dårlig løsning: bare se på effekten innad i et subsett
  + Bedre løsning

Tilbake til økologisk mat

* Oransje: folk som spiser 20 økologiske grønnsaker og trener fra 4-14 timer
* Blå: folk som trener 10 timer i måneden og spiser fra 14-26 økologiske grønnsaker
* Begge tilfeller ser vi hva modellen predikerer når den andre variabelen er konstant

Suppressorer

* I alle eksemplene vi har sett så langt blir effekten av en uavhengig variabel mindre når jeg kontrollerer for en confounder.
* Noen ganger er det lite/ingen assosiasjon mellom to variabler inntil du kontrollerer for en tredje.
  + Den tredje variabelen kalles da en supressor

Avdekking av supressoreffekt

* Over: i dette tenkte eksempelet mer enn dobler størrelsen på sammenhengen mellom fritid og lykke seg når man legger til inntekt.

Suppressorvariabler

* De uavhengige variablene er negativt korrelert, men positivt forbundet med lykke.
* De uavhengige variablene er positivt korrelerte, men har motsatt effekt på den avhengige.
* Nuller hverandre ut

Standardiserte regresjonskoeffisienter

Sammenligning av stigningskoeffisienter

* I modellen under er (høyere) utdanning målt i år, og inntekt målt i 1000 nok. Hvilken uavhengig variabel er sterkest assosiert med lykke.
* Det gir som regel ingen mening p sammenligne de ustandardiserte koeffisientene, frdi variablene kan ha veldig ulik skala.
* For å kunne sammenligne regresjonskoeffisienter må vi sette dem på en liknende skal, der «en enhet økning» er en sammenlignbar størrelse.

#install.packages(‘’lm.beta’’)

library(lm.beta)

mod2 🡨 …

Partiell korrelasjon

* Korrelasjonen mellom to variable der den effekten som overlapper med en eller flere andre variabler er fjernet.
* Det er rimelig deler av korrelasjonen mellom lykke og størrelsen på leiligheten skyldes inntekt. Hvordan kan jeg kontrollere for inntekt?
* Vi vet at bakenforliggende variabler kan påvirke, og ved å regne ut partiell korrelasjon, så tar man vekk systematisk sammenheng mellom målet på de to variablene.
* Når du mistenker en bakenforliggende faktor, så regner man ut

#install.packages(‘’ppcor’’)

Library(ppcor)

#pcor.test(…

Polynomisk regresjon

Ikke-lineære forhold og polynomisk regresjon

* Lineær regresjon er mer fleksibel enn man skulle tro, og kan modellere ikke-lineære relasjoner også.
* En utbredt måte å regresjon til å se på ikke lineære sammenhenger er ved å modellere polynomisk regresjonsfunksjon.

summary(lm(lykke ~ Alder + I(Alder^2)))

Innflytelse

* Antagelser i regresjon
* Uteliggere
* Residualer
* Leverage
* Influens
  + Cook’s D

# 4 Inferens og regresjon

Ekstroversjon og intelligens

* Prøve å utlede: hvordan trekke slutninger basert på data fra populasjonen

Hva er slutningsstatistikk?

* I første forelesningen tok vi for oss deskriptiv statistikk
* - verktøy for å kvantifisere vitkige egenskaper til et datasett
* Som navnet antyder er dette beskrivelser av utvalger og vi trekket ikke sllutninger om hva den ‘sanne’ verdien i befolkningen er.
* Kan det hende at du i en regresjonsanalyse får bhat =/= 0 selv om variablene egt ikkeer assosierte?
  + Kan sammenhenger komme til uttrykk ved tilfeldighet i noen …

Hva dere bør få med dere av inferens?

* Forstå hva vi mener med begrepene samplingfordeling og standardfeil, og forstå hvordan disse brukes i statistisk hypotesetesting
* Burde være i stand til å regne ut\_
* Standardfeilen til bhat\_1
* T-statistikken som brukes til å teste om individuelle regresjonskoeffisienter er statistisk signifikant.
* F-statistikken, som brukes til å teste den samlede innflytelsen til flere uavhengige variabler
* Et konfidensintervall for bhat\_1

Populasjon og utvalg

* Utvalg: et sett med observasjoner som er trukket fra en større populasjon.
* Sampling: det å trekke ut et subset fra populasjonen.
* En statistikk: enhver egenskap (verdi) regnet ut på et utvalg (altså ikke på hele populajsonen).
* Merk: når vi trekker…

Hypotesetesting

* Innen statistikk er en hypotese en påstand om en populasjon atså en parameter ved populasjonen har en spesifikk verdi. Typisk har vi to rivaliserende hypoteser.
  + Nullhypotesen
  + Forskningshypotesen

Samplingfordeling ved simulering

* Samplingfordelingen til en statistikk er fordelingen av de verdiene statistikken tar på tvers av ulike utvalg
  + Standardavviket til en samplingfordeling kalles standardfeilen (SE)
  + Frdelingen over kalles…

Bruk av samplingfordelingen

* Sentralgrenseteoremet (central limit theorem): summen av uavhengige og identisk fordelte variabler er næremer seg en normalfordelt ettersom utvalgsstørrelsen vokser.
  + Dette er edet viktigste teoremet innen statistikk, og basisesn for nesten all klassisk slutningsstatistikk.
  + Moralen er : noen ganger vet vi hva samplingfordelingen til vår statistikk er.

Hypotesetesting oppsummert

1. formuler nullhypotese, og en alternativ hypotese
2. finn en teststatistikk med kjent samplingfordeling under H0
3. samplingfordelingen brukes til å finne p-verdien, definert som; sannsynligheten for å få en…

Litt om frihetsgrader

* df til en statistikk antall verdier i statistikken «får lov til» å variere.
* Antall frihetsgrader vil være gitt som; antall observerte verdier – antall estimerte parameter i en statistikk.
* I varians er det bare (n-1) observasjoner som brukes til utregningen, siden den siste «låses» av den estimerte verdien for gjennomsnittet.

Kvadratsummer med frihetsgrader

* TSS = SSM + SSE
* df\_TSS = df\_SSM + df\_SSE

hva er utvalgsstørrelsen

* ved å bruke anova() så kan man få sum of squares-tabell
* TSS: 100 + 200 = 300
* p: antall uavhengige variabler (her: 1)
* n: 1+98=99. 99 er n-1, så n må være 100

Standardfeilen til B\_1 (stigningstallet)

* (SE( hat(b)\_1
* Er et estimat på variabiliteten (standardavviktet)
* Residual standard error (conditional standard deviation) et et estimate av i uttrykket.

Standard error

t-statistikken: testing av individuelle regresjonskoeffisienter

t-fordelingen

* Formen på en t-fordeling bestemmes av ett parameter, som kalles frihetsgrader (df)
* De blå kurvene over er t-fordelinger med 1 (mørkeblå) til 16 (lyseblå) frihetsgrader.
* Legg merke til at når antall frihetsgrader øker, så …

Kritiske verdier

* Når nullhypotesen er sann (b\_1=0), så følger t-statistikken i ulike tilfeldige utvalg på størrelse 1o en t-fordeling med 8 frihetsgrader.
* Jeg kan da bruke en tabell for å finne de kritiske verdiene til t-statistikken.
* Under H0 er t-statistikken t-fordelt med (n-p-1) frihetsgrader.

Bruk av t-tabellen for å finne kritiske verdier

* Nå kan man gjøre det i R.
* With fewer degrees of freedom there is more uncertainty
* t= bhat\_1 / SEhat(bhat\_1) =0.728/0.273
* dele error x på estimate x (bhat\_1)

oppsummert av t-testen

* samplingfordeling under H0: under H0 følger t-statistikken en t(n-p-1) fordeling.

Hvilken uavhengig variabel har lavest p-verdi?

* Stor verdi på t-verdien = liten verdi på p-verdien

Konfidensintervaller (CIs)

* En annen måte å uttrykke usikkerheten i et estimat
* Enhver verdi innad i et 95% konfidensintervall utgjør en mulig hypotese om parameterets verdi som du ikke kan forkaste på et 0.05 nivå. Enhver verdi utenfor intervallet kan forkastes på et 0.05 nivå.
* Konfidensintervall gir mer informasjon enn en p-verdi.
  + Gir samme info som p-verdien + mer
  + I tillegg kan vi se hvilke andre verdier vi kan forkaste.

Utregning av konfidensintervaller

* confint()

# 5 Kategoriske predikatorer, interaksjon, mediering

Dagens forelesning

1. avsluttende om F-test og konfidensintervaller
2. Kollinearitet
3. kategoriske uavhengige variabler
4. modellering av interaksjon
5. mediering
6. modellbygging/modellvalg

oppsummering av t-testen

* Hypotese
  + H0
  + H1
* Test-statistikken: stigningtallet / standardfeilen
* Samplingfordelingen under H0:
  + Under H0 føger t-statistikken en t(n-p-1) fordeling.

Konfidensintervaller

f-testen: testing av flere regresjonskoeffisienter samlet

oppsummering av f-testen

* Hypotesene
  + H0
  + H1
* Test-statistikken
  + F = MSM / MSE
* Samplingfordeling under H0
  + Under H0 følger F-statistikken en F(p, n-p-1) fordeling.

Samplingfordelingen til f-statistikken

* Bestemmes av de to tallene til frihetsgradene (df)
* Vil se veldig forskjellige ut avhengig av hvor mange frihetsgrader

Kollinearitet

Underlige resultater

* Lave p-verdier hver for seg, ingen er signifikante sammen

Kollinearitet

* Graden av lineær sammenheng mellom flere forklaringsvariabler i en multippel regresjonsmodell.
* Problem som oppstår når summen av variabler begynner å korrelere veldig mye
* Blir denne for høy (korrelasjon >>0.8) kan det medføre problemer
* Usikkerheten rundt koeffisentene øker og det bir vanskelig å se hvilke variabler man kan stole på.
* Konsekvenser av kollineratet
  + B’ene er ikke til å stole på grunnet stør økning i standardfeilene.
  + Vanskelig å vurdere hvilken av de uavhengige variablene er av betydning.
  + Justert R2 kan begynne å synke.
  + Tilpasning av modellen er ikke påvirket.

Kollinearitetsdiagnostikk

* Se etter store korrelasjoner mellom de uavhengige variablene (>0.8).
* VIF (variance

Kategoriske variabler

* Kan inkl. kategoriske variabler, men du må rekode dem. kan gjøres automatisk

Dummyvariabler

* En dummyvariabel (indikatorvariabel) ta bare verdiene 0 eller 1.
* Det er kanskje ikke meningsfullt å snakke om «én enhets økning i MANN», men siden variablene bare har to nivåer kan vi si at b1 er
  + b1 vil fange opp differansen mellom to variabler

Kategoriske uavhengige variabler (faktorer)

* by er en kategorisk variabel med flere nivåer, og det gir ikke mening å snakke om betydningen av en økning på en enhet.
* For å inkludere kategoriske variabler som uavhengige variabler i regresjonsmodellene

F-test

* Hypotesen om at det er samme gjennomsnitt i alle grupper kan vurderes med en F-test
* H0: b2=b3=0

Betydning av referansekategori

* Annen måte å kode variablene på
* Vil ikke påvirke tilpasningen av modellen
* Tolkningen vil være forskjellige

Interaksjon (moderasjon)

* Effekter er additive dersom den kombinerte effekten av et lik summen av enkelteffektene.
* Interaksjon i statistikk er når effekten av to variabler ikke er additiv, men når styrken på forholdet mellom to variabler avhenger av en tredje variabel.

Kryssproduktledd

* Regresjon som verktøy for å se på interaksjon

Interaksjon i regresjonsmodeller

* Interaksjon modelleres vanligvis med et kryssproduktledd
* Er interaksjonsleddet statistisk signifikant?
* Differansen mellom p-verdiene

Mediering

* Mediator er en variabel som formidler deler av effekten/risikoen av andre uavhengige variabler på utfallet.
  + På rent statistisk grunnlag kan du ikke skille mellom en mediator og en konfunderende variabel. Dette må gjøres på teoretisk grunnlag.

Klassisk måte å påvise mediator

* Trinn 1: vis at den første variabelen (ekstroversjon) er assosiert med den avhengige [sti c].
* Trinn 2: vis at den første variabelen er assosiert med mediatoren [sti a].
* Trinn 3: vis at mediatoren er assosiert med utfallet [sti b].
* Trinn 4: for å kunne hevde at effekten av ekstroversjon på lykke medieres av antall venner, bør effekten av ekstroversjon blir vesentlig redusert når venner også inkluderes i modellen.

Oppsummering av 3-variabel assosiasjoner

* Multiple causes
* Spurious X1 Y association
* Chain relationship (complete mediation)
* Both direct and indirect effects of X1 on Y (partial mediation)
* Interaction

Hvordan velge endelig modell?

* Hvorvidt en uavhengig variabel vil være signifikant, avhenger også av de andre uavhengige variablene i modellen (dersom de er korrelerte). Hvordan skal du gå frem for å bestemme hvilke uavhengige variabler…

1. Hierarkisk regresjon
2. Skrittvis (stepwise)
3. Informasjonskriterier

Metode 1: hierarkisk regresjon

* Grupper av uavhengie variable blir trinnvis lagt til.
* F-testen kan benyttes til å vurdere om et trinn signifikant forbedrer modellen.

Metode 2: skrittvis (stepwise)

* Skrittvis regresjon er en algoritme som automatisk velger ut variabler for en regresjonsmodell.
  + Forward selection: start uten uavhegige variabler i modellen. Test hvor mye hver enkelt mye variabel ville forbedret modellen, og legg til den som forbedrer modellen mest. Gjenta denne prosessen til ingen nye variabler signifikant forbedrer modellen.

Begrensninger med automatiske fremgangsmåter

* Stepwise regresjon er en fullstendig automatisert prosess der dataene får…

Over vs. undertilpasning (Scylla og Charybdis)

* Overtilpasning: dårlig prediksjon som følger av at du lærer for mye fra data.
* Undertilpasning: dårlig prediksjon som følger av at du lærer for lite fra data.
* Du kan alltid på bedre tilpasning ved å inkludere flere prediktorer.
* Alle datasett inneholder både systematiske og usystematiske effekter (støy), så komplekse modeller kan predikere mindre varians i nye datasett.

Hva skjer når modellen blir for kompleks?

* Over: den grønne kurven er ‘sann’. Den røde passer best, men tilpasser seg mye av støyene i dataene.

k-fold

* Kryssvalidering innebærer å dele datasettet ditt inn i et treningssett og et testsett. Så bygger du modellen utifra treningssettet, og gjør den endelige vurderingen av modellen etter hvor godt den presterer i testsettet.
* Kryssvalidering er hjørnesteinen i moderne maskinlæringsmodeller.
* Merk:

1. det er lett å forklare varians, det er mye vanskeligere å predikere nye data.

2. først øker forklart varians i testutvalget, men ettersom modellen blir mer og mer kompleks, begynner ytelsen i testdataene å synke.

Metode 3: informasjonskriterier (AIC/BIC)

* En klasse statistikker som er utviklet for å optimalt balansere mellom over og undertilpasning, tilpasning og kompleksitet.
* Modellen som velges er den som har lavest verdi på informasjonskriteriet, her akaike information criterion (AIC).

# 6 Kji-kvadratanalyse

Kategorisk data

* Data der verdiene utgjør et bestemt antall distinkte kategorier
  + Nominelle variabler: ingen rangering av kategoriene
  + Ordinale variabler: distinkte kategorier, men de kar en rangering
* Ikke mulig å bruke regresjonsanalyse på kategorisk data
* Når den avhengig verdien er kvantitativ, så er regresjonsanalyse egnet
* Hva om man ønsker å se på sammenhengen mellom to kategoriske variabler?

Grunnleggende sannsynlighetsregning/regler

Stokastisk (random)

* Et fenomen er stokastisk dersom individuelle utfall ikke er gitt, men dersom utfall følger en bestemt fordeling over gjentatte repetisjoner.
* Utfallsrom: settet av alle mulige utfall av et stokastisk fenomen.
* En hendelse er et bestemt subsett av utfallsrommet.
  + I statistisk litteratur refereres disse ofte til med en bokstav.

Venn diagram

* Et venn-diagram viser alle mulige logiske relasjoner mellom samlinger av forskjellige sett.
* Brukes ofte som en uformell representasjon, og eksakte størrelser/former av regioner er ikke viktige.

Sannsynlighet

* Dersom alle utkommene av et stokastisk fenomen forekommer like ofte, definerer vi sannsynligheten for utfallet A som:

Additativ regel

* Hvis A og B er to hendelser, så er sannsynligheten:
* For gjensidig utelukkende sett forenkles dette til:
  + P…

Simultan sannsynlighet

Betinget sannsynlighet

* For ethvert to utfall A og B, der P(B) ikke er lik 0, er den betingede sannsynligheten av A gitt B definert som:

Uavhengighet

* To utfall er uavhengige dersom:

Multiplikasjoneregelen

* Dersom A og B er uavhengige utfall, så er den simultane sannsynligheten (sannsynlighet for begge) gitt ved….

Marginal sannsynlighet

* The relationship between parenthood and seat belt use
* Totaelene i en tabell

Simultan sannsynlighet

* Verdiene i de forskjellige kategoriene

Hva har vi lært av setebeltebruk?

* De observerte frekvensene for bruk av setebelte/ha barn i bilen later til å være forskjellige fra de vi ville forvente dersom variablene var uavhengige.

Kji-kvadrat tester

Vi vil se på to anvendelser:

1. kji-kvadrat goodness of fit test
   * Teste i hvilken grad
2. kvi-kvadrat test av uavhengighet

Kji-kvadrat statistikken

* Avstandsmål

1. Goodness-of-fit test

* Lar oss sammenligne de observerte frekvensene med de vi ville forventet dersom de var trukket fra en bestemt fordeling.
  + her ønsker jeg å sammenligne de ovserverte frekvensene med de vi ville forventet om deltagerne bare gjettet.
  + H0: de observerte dataene skuldes ren gjetting
  + H1: de observerte dataene skyldes ikke ren gjetting
* Under H0 følger den observerte verdien for kji-kvadrat…

Chi-kvadrat tabell

2. kji-kvadrat test av uavhengighet

* er kjønn og politisk tilhørighet uavhengige variabler?

Analyse av residualene

* sammenligning av de observerte og forventede frekvensene lar oss teste hypotesen om uavhengighet.
* Forskjellen f0-fe, differansen mellom observert frekvens og forventet frekvens kalles et residual.
* Hvor stort må…

Avsluttende om kji-kvadrat analysene

* Bergenset innsikt følger av å forkaste H0.
  + Selv om du kan forkaste hypotesen om uavhengighet, sier kji-kvadrat tester sier i seg selv ikke noe om typen eller styrken på sammenhengen.
  + Testen er ikke egnet for små utvalg
  + For valide resultater kreves …

# 7 Variansanalyse

Målenivåer

* Ting vi måler inn i fire ulike kategorier:
  + Nominal: kan ikke rangeres etter størrelse
  + Ordinal: kan rangeres etter størrelse, men avstanden er ikke meningsfull.
  + Intervall: avstanden mellom verdien er meningsfull
  + Forholds/ratio: mengdeskalaen har også et nullpunkt
* De tre siste er kvantitative variabler. Nominale variabler er kvalitative

Regresjon med dummyvariabler

* E.g., tre barn ble utsatt for tre ulike intervensjoner, som skulle øke prestasjonen på en lesetest.

Tabulering av dataene

* Vi kan oppsummere noen tall for hver gruppe

Dummyvariabler i R

* Som dere også har sett tidligere, en dummyvariabel er en variabel som kun tar verdiene 0 og 1. for å representere tre ulike grupper trenger vi to dummyvariabler.
* E lager dummyvariabler automatisk for oss.
* Call:
* lm(formula = Prestasjon ~ Gruppe, data = lesenivå)
* Gruppe A er referansenivå, mens Gruppeb og Gruppec representerer forskjellen mellom disse gruppene og gruppe A.
* R gjør dette med dummyvariabler automatisk, og begynner med A
* La z1 og z2 være dummyvariablene. Disse tilsvarer Gruppeb og Gruppec i R-utskriften.
* R lager automatisk så mange dummy-variabler som trengs med 0 og 1.
* Obs: R skjønner ikke uten videre at tall betyr grupper!
  + R tror Gruppe er kvantitativ
  + Vi må passe på at gruppevariabelen er en faktor:
    - Lesedata$Gruppe <- factor(lesedata$Gruppe)
  + Konvertering til factor ga oss dummyvariabler igjen

Regresjon med dummyvariabler

* Regresjonsmodellen med to grupper er
  + where alpha er intercept/skjøringspunkt, beta1 og beta2 er effekten av de to dummy-variablemene, og e er residualet (støyledd).
    - Gjennomsnitt i gruppe a er alpha
    - Gjennomsnitt i gruppe b er alpha + beta1
    - Gjennomsnitt i gruppe c er alpha + beta2
  + Dette gir oss en tolkning av regresjonskoeffisientene som forskjellen mellom grupper:
    - Beta1 = mub – mua
    - Beta2 = muc -mua

Er forskjellen signifikant?

* Nullhypotesen er at det ikke er forskjell mellom gruppene,
  + H0: mua =mub = muc
* H0 er det samme som beta1 = beta2 = 0. alternativhypotesen er at det er forskjell mellom minst to av gruppene.
* Vi kan bruke anova-funksjonen i R til dette:
  + Anova(lesemodell)
* Alternativt kan vi bruke aov-funksjonen til å gjøre variansanalysen direkte
  + Summary(aov(Prestasjon ~ Gruppe, data = lesedata))
    - Direkte global test

Testing av parvise gjennomsnitt

* Testen vi har gjort sjekker om det er en forskjell mellom noen av gruppene, men ikke mellom hvilke. Vi kan sjekke dette også.
* La være gjennomsnitt i gruppe g, være antall observasjoner i gruppe g, og være Residual Mean Square. Da er et konfidensintervall for forskjellen…
* Stor residual mean square har du en høy usikkerhet. Om av residual mean square, så er ganske sikkert at gruppemedlemsskap forklarer mye.
* Kan vi finne et konfidensintervall for forskjellen mellom gruppe c og gruppe b, muc – mub?
  + Lesedata %>%
    - Får ikke direkte ut svaret
  + Gjennomsnittlig kvadrat til forskjellen (feilen) mean square
* t-en kommer fra t-fordelingen med antall frihetsgrader gitt ved differanse mellom antall observasjoner og antall grupper, df = 9-3. vi velger 95% signifikansnivå, og får derfor alpha= 0.025, siden vi skal ha 2,5% i hver hale av fordelingen.
  + C(qt(p=.025, df=6), qt(p.975, df=6))
  + Intervallet inneholder ikke null, så vi konkluderer at gjennomsnittet i gruppe c er lavere enn i gruppe b.

Mutlippel testing

* Når antallet grupper vokser, vokser antallet parvise sammenligninger mye raskere.
* Hvert 95% konfidensintervall har 5% sjanse for å ikke inneholde den sanne verdien. Hvis vi f.eks. sammenligner 7 grupper, har vi konfidensintervall, og kan forvente at 21x0.05=1.05 av disse er feil.

Bonferroni-korreksjon

* Endrer bredden på konfidensintervallene, slik at vi kontrollerer den totale sannsynligheten for å gjøre feil.
* Metoden er enkel å bruke. For eksempelet tidligere, dersom vi sammenligner alle tre par, så deler vi på 3. hvert enkelt konfidensintervall har derfor t-verdi basert på 0.025/3=0.00833.
  + qt(p=.025/3, df=6)
* intervallet vårt er ikke signifikant lenger.
* Tar hensyn til å gjøre mange tester og gjøre flere feil
* Korreksjonen gjør intervallene bredere slik at det er vanskeligere å gjøre feil.
* I praksis er bonferroni-korreksjonen for streng, den går ofte under 5%. en konservativ test, men man mister styrke for å finne det man er ute etter

Tukeys metode for multippel korrelasjon

* Bonferronis metode er konservativ. Den garaterer at den totale feil er mindre eller lik signifikansnivået. Tukeys metode er litt mer liberal, ved at den sikrer at den forventede sannsynligheten for feil er lik signifikansnivået. Det finnes også en enkel funksjon for dette i R:
  + Anova\_modell <- aov(Prestasjon ~ Gruppe, data = lesedata)
  + TukeyHSD(anova\_modell)
  + Trenger bare å teste 3 par, ikke 6
* plot(TykeyHSD(anova\_modell))
* bare relevant om man faktisk sjekker alle

Litt mer variansanalyse

* For g grupper kan vi teste nullhypotesen
  + H0: mu1 = mu2 = …=mug
* Mot…
* Jo større forskjell det er mellom gruppene enn innad i gruppene, jo mer sikker er du på at det er …

Variansanalyse

* Teststatistikken er
* F = varians mellom / varians innad
* Frihetsgradene er df1 gitt ved antall grupper minus én, og df2 gitt ved antall observasjoner minus antall grupper.
* Tolkning av F-statistikken: stor varians mellom …
* ToothGrowt$dose <- factor(ToothGrowth$dose)
* Summary(aov(len ~dose, data = Totthgrowth))
* Kolonnen mean sq viser varians, raden Residuals tilsvarer within-group og group tilsvarer between-group

Toveis variansanalyse

* Vi har to uavhengige variabler. Vi kan også utføre treveis, fireveis, osv. men da blir det fort vanskelig å skjønne hva vi holder på med.
  + boxplot(len ~ dose + supp, data = ToothGrowth)
* vi har nå to nullhypoteser, én for hver faktor. For dose er nullhypotesen at det ikke er forskjell i lengden…
* summary(aov(len ~ dose, data = ToothGrowth))
* summary(aov(len ~ dose + supp, data = ToothGrowth))
* ved kji-kvadrat-analyse som begynner i 0, tar man n-1 fordi man er interessert i den høyre halen
* grense for å forkaste H0
* qf istedenfor pf, så får man quantile

toveis variansanalyse som multippel regresjon

* modellen vi her har estimert er nøyaktig den samme som
  + mod2v <- lm(len ~ dose + supp, data = ToothGrowth))

Interaksjoner

* Er effekten av dose avhengig av hvordan C-vitaminene leveres? Er effekten a leveringsmetoden avhengig av dosen?
* Boxplot(len ~dose \* supp, data = ToothGrowth)
* Anova(intrxmod)
* Teste interaksjon: \* istedenfor +

# 8 Mer variansanalyse og bootstrapping

Interaksjoner

* Når vi har to eller flere kategoriske variabler i en variansanalyse
* Ser på hvordan ideologi avhenger av partitilknytning
* Avhengig variabel, ideologi, på skala: 1-7, veldig liberal til veldig konservative
* Ideologi avhenger av partitilhørighet, men avhenger det også av kjønn?

Interaksjoner

* Steg 1, toveis anova med interaksjoner.
  + Dummyvariabler for parti, p1=1 og p2=1 for uavhengige
  + Dummyvariabel for kjønn: s=1 for kvinner
* Regresjonsmodell
  + E(y)= alpha + beta1p1+beta2p2+beta3s+beta4p1s+beta5p2s
  + Impliserer, ved å plugge inn riktige verdier for dummyvariabler:
    - Republikanere: E(Y)=alpha+beta3 for kvinner og E(y)=alpha for menn.
    - Uavhengige: E(y)=alpha+beta2+beta3+beta5 for kvinner og E(y)=alpha+beta2 for menn.
    - Demokrater: E(y)=alpha+beta1+beta3+beta4 for kvinner og E(y)=alpha+beta1 for menn.
* R setter opp dummyvariablene for oss automatisk:
  + mod <- lm(ideologi~partyid\*sex, data=party\_data)
  + Siden kvinner har -0.128 så vil kvinners skåre være litt lavere enn menns
  + Interaksjonene er ikke signifikante
  + Vi har ikke noe belegg for å si at hypotesen vår ikke stemmer, kjønn har ikke noe å si

Toveis variansanalyse uten interaksjoner

* I modellen tillater vi at effekten av partitilhørighet på ideologi varierer mellom kjønn. Eller tilsvarende, at effekten varierer mellom partitilhørigheter. En modell uten…

Skal vi ha med interaksjonen?

* Modell uten interaksjon:
  + mod0 <- lm(ideologi~partyid+sex, data=party\_data)
  + Sammenligning:
    - anova(mod, mod0)
* I mod har vi brukt opp to frihetsgrader på interasjonen
* Modellen mod er ikke signifikant bedre enn mod0, så dermed kan vi bruke mod0 som ikke har interaksjon. Vi går for den mindre kompliserte modellen.

Noen huskeregler for interaksjoner

* Dersom interaksjonen er signifikant, men hovedeffektene ikke er det, skal hovedeffektene likevel med i modellen. Interaksjoner gir ikke mening uten at hovedeffekten også er er.
* Interaksjoner betyr at effekten av én variabel basert på verdien av en annen variabel. Begge må derfor tolkes samtidig. For eksempel, dersom interaksjonen hadde vært signifikant i vår modell for ideologi:
* ---

Variansanalyse med repeterte målinger

* Head(Anorexia)

Repeterte målinger

* To målinger for tre behandlingsgrupper: kontroll c, kognitiv terapi b, og familieterapi f.
* Naturlig å sette kontrollgruppa til referansenivået
* La s=1 i tidspunkt 2 og 0 i tidspunkt 1
* La t1=1 for Treat= «b» og t2=1 for treat= «f».
* Følgende virker naturlig:
  + E(y)=alpha+beta1s+beta2t1+beta3t2+beta4st1+beta5st2
  + Alpha er intercept
  + Beta1 er endring i kontrollgruppa(
* Problem: disse observasjonene er ikke uavhengige. Vi har to stykk fra hver person.
  + Lage random intercept som beskriver hver person
  + Repeterte målinger er ikke helt tilfeldig/uavhengig siden det er den samme personen som testes, og dermed må resultatene ligge i nærheten av hverandre. Skiller mellom variasjonen skiller seg fra mine data, og hvordan mine data skiller seg fra gjennomsnittet.
* Kilder til varians: variasjon mellom individer og innad i individer.
  + Mest variasjon innad: lite systematisk mellom menneskene

Random effects

* Vi legger til en dummy-variabek for hver eneste deltager
* Koeffisientene kalles random intercepts. De fanger opp den systematiske variasjonen mellom individer. Det vanlige residualleddet fanger opp variasjonen innad i individer.
* Random effects er ikke parametre…

Random effects

* Subset(Anorexia, subject=23)
* Det er mulig å bruke aov()-funksjonen i R. Appendix A i boka. Her illusterer jeg funksjonen lmer() fra pakka lme4.
  + library(lme4)
  + mod <- lmer(Weight ~Timepoint \* Treat + (1|subject), data=Anorexia)
  + summary(mod, correlation=FALSE)

varians innad og mellom

* etter vi har tatt hensyn til de uavhengige variablene, er det større variasjon innad enn mellom.
* For en ny person i populasjonen:
  + E(y)= alpha+beta1s+beta2t1+beta3t2+beta4st1+beta5st2
* Effekten av kognitiv terapi er:
  + (alpha+beta1+beta2+beta4)-(alpha+beta2)=beta1+beta4

Konfidensintervaller og p-verdier

* Når modellene blir kompliserte, blir det vanskelig å finne nøyaktige formler. Vi trenger da å bruke approksimasjoner. Vi kan få p-verdiene for betaene, men ikke så lett for differansen mellom betaer.
* Og vi kan få konfidensintervallene av betaene:
  + confint(mod)

bootstrapping

* hvordan får vi konfidensintervaller for effekten av kognitiv terapi beta4+beta1 eller effekten av familieterapi beta5+beta1?
* her finnes det en metode som alltid virker.
* En simuleringsmetode, som fungerer som følgende:
  + Bestem et stort tall T, f.eks. 10000
  + Trekk radene fra det opprinnelige datasettet, med tilbakelegging, slik at du får T nye like store datasett.
  + Tilpass T modeller, én til hvert av de nye datasettene, og for hver modell lagre verdien for statistikken for intervallet…

Bootstrapping steg for steg

* Finn ut hvor mange observasjoner vi har:
  + (n <- nrow(Anorexia))
* Trekk n nye rader og ta en titt på dem:
  + nye\_rader <- sort(sample(n, n, replace=TRUE))
  + nye\_rader
* lag et nytt datasett baser på nye\_rader:
  + bootdata <- Anorexia[nye rader, ]
  + head(bootdata)
* vi tilpasser modell til dette datasettet:
  + bootmod <- lmer(Weight ~Timepoint \* Treat + (1|Subject), data=bootdata, REML=FASE)
* og lager beta 4+beta1
  + fixef(bootmod)
  + fixef(bootmod)[[“Timepoint2”]]+ fixef(bootmod)[[“Timepoint2:Treatb”]]
* vi kan sette det hele sammen
  + boot\_beta1beta4 <- rerun(1000, {

nye rader <- sample(n, n, replace=TRUE)

bootdata <- Anorexia[nye\_rader, ]

bootmod <- lmer(Weight~Timepoint \* Treat + (1|Subject),

data=bootdata)

…

* + hvor stor andel av diagrammet vårt ligger på den negative siden?
* 95% konfidensintervall for beta1+beta4 er
  + Quantile(…
* P-verdi for H0: beta1 + beta4 \_< 0 mot alternativet beta1+beta4 > 0 er:
  + Mean(boot\_beta1beta4 < 0)
* P-verdien er ensidig mens konfidensintervallet er tosidig.

Bootstrapping

* Nyttig når man trenger konfidensintervaller og når disse ikke kommer rett ut av programmet.
* R-pakken boot kan gjøre jobben i mange tilfeller.
* For mixed models som vi brukte her, …

Oppsummering

* Interaksjoner
  + Når vi har to eller flere faktorer i en variansanalyse må vi sjekke om det er interaksjoner mellom dem.
* Repeterte målinger
  + Når vi har to eller flere repeterte målinger, må vi ta hensyn til at målingene av samme individ er korrelert.
  + Random effects lar oss skille mellom variasjonen mellom individer og variasjonen innad i individer.
* Bootstrapping
  + Simuleringsmetode som gir konfidensintervaller. Nyttig når programmet ikke gir deg akkurat det du er ute etter.

# Siste forelesning

Vi har vært gjennom tre temaer:

* Regresjon
* Kji-kvadratanalyse
* Variansanalyse

Brukte uttrykk for å regne stigningstallet:

* Hat: estimat
* B^hat0: skjæringspunkt
* B^hat1: stigningstall. Kovarians/varians
  + Variansen til en variabel er ofte referert til med en liten s

Oppg.1

* kovarians: 17.06
* standardavviket til søvn: 13.27

1. b^hat1 = s\_xy / s^2 = 17.06 / 13,27^2 = 0,097

* forstå hvordan verdiene skal tolkes

1. V^hat= b^hat0 + b^hat1 SØVN.C

* 67,8 + 0,097\*SØVN.C
* 67,8 + 0,097\*100 = 77,506
* Sentrert: mest vanlig å sentrere ved å trekke fra snittet

Oppg. 2

* Modell 2 inneholder følgende uavhengige variabler; SØVN.C, ABSTRAKT og SØNVN.CxABSTRAKT
* Interaksjonseffekter vi leter etter
* Er det de som har måttet sove mer enn snittet husker mer? Er det en sammenheng mellom søvn og abstrakte vs. konkrete ord?
* Hva slags ord er det som huskes best jevnt over? Konkrete
* Effekten for abstrakte og konkrete ord er ikke den samme
* V^hat= b^hat0+b^hat1 S + b^hat2 A + b^hat3 S\*A
* Abstrakt ord-lesere ligger 14,8 enheter under konkret ord-lesere
* Man kan ikke si noe om effekten av søvn uavhengig av gruppe
* Effekten er sterkere for de som leser abstrakte vs. konkrete ord
* Er interaksjonseffekten signifikant
* Kontrollert for alder, så er effekten av søvn mindre. Søvn er en konfunderende variabel.
* Vi bruker multippel regresjon for å kontrollere for dette
* Regresjonsplanet er et hjelpemiddel
* Endring i søvn kan forstås som konfunderende effekt av variabelen alder
* Hva hadde vi kalt alder dersom effekten av søvn blir mye større? Suppressorvariabel
* Mest interessert i effekten av alder på søvn. Søvn er mindre signifikant når man kontrollerer for alder.
* Inferens: standardkoeffisienten til variabelen
* Ønsker å se en liten som mulig standardfeil for da er man mer sikker på estimatet man har gjort
* Hvordan legge opp studien sånn at man får så liten standardfeil som mulig
* Liten range i søvn vil gi større standardfeil

Signifikansverdier

* I modell 4 legges følgende variabler til: FYSISK, SOSIAL og HODEOMKRETS
* P-verdien er mest direkte og enkle indikator å se på
  + Om man ikke kan se p-verdien, så kan man se på t-verdien. Da kan man slå opp i en tabell og se på den kritiske verdien. Men man kan også i og med at det er et såpass stort utvalg, så vet vi at vi leter etter verdier som er mindre enn -2 og større enn 2. det er de kritske verdiene om den var z-fordelt (standard normalfordeling). Modellen ahr veldigmange frihetsgrader, og når vi har nok frihetsgrader bruker vi +-2.
  + Når vi ikke har t-verdien kan vi regne ut t-verdien med stigningskoeffisient/standardfeil. 3,6 er signifikant

Konfidensintervall

* Har nedre og øvre grense
* Om den nedre grensen er over 0, så kan ikke 0 ligge i intervallet, så dermed er faktoren signifikant
* Om intervall inneholder 0: ikke signifikant
* Om intervall ikke inneholder 0: signifikant
* De ustandardiserte koeffisientene er relativt identiske
* De standardiserte er veldig forskjellig
  + Hvorfor?
    - Én enhets endring i FYSISK er ett minutt økning i fysisk aktivitet, enheten er liten
* Ustandardisert vs. standardisert
* Standardisert gjør faktorer sammenlignbare
* Modellen kommer med en del antagelser:
  + De dreier seg om epsilon-leddet
  + N står for normalfordelt
  + ~ betyr ‘er fordelt sånn’
  + Epsilon er trukket fra en normalfordeling
  + Gjennomsnittet er 0
  + Trukket fra fordeling med den samme variansen, det er ikke noe i i symbolet. Spredningen rundt alle observasjoner skal være den samme.
  + Når kunne residualene være korrelert? Innad i forskjellige grupper er mer korrelert enn mellom. Alle observasjoner er ikke uavhengige hverandre.

Ekstreme residualer

* Problem: kan få stor konsekvens for regresjonskoeffisientene
* Vi minimerer disse SSE. Vi finner denne regrsjonslinjen som gjør at regresjonslinjen blir mest mulig.
* Sum of squares between: n-1
* F-value: mean square value/mean square residuals

Tolke toveis-variansanalyse

* F-verdien er noe mindre
* Effekten av person er nesten lik effekten av speil i rommet
* Leter etter verdier mindre enn -2 eller større enn 2.