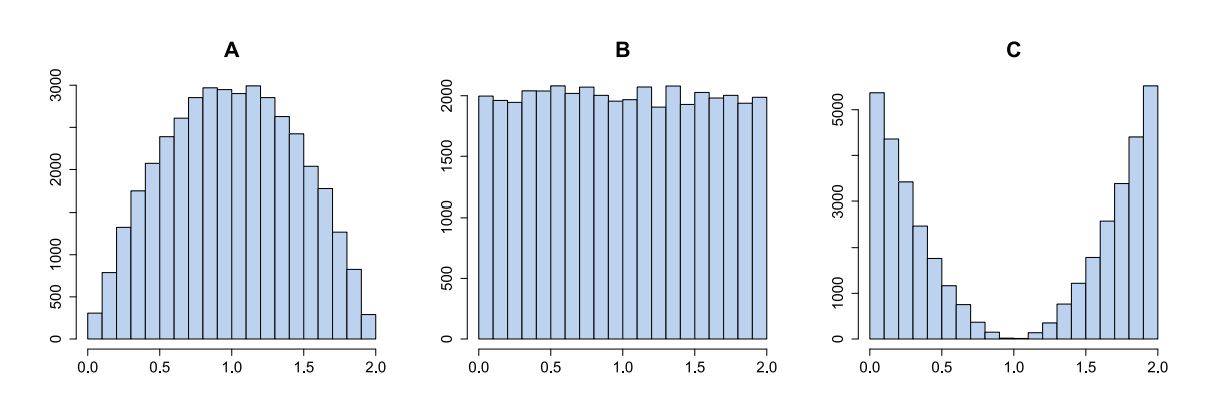
Eksamenstakeaway

# Sandwichliste

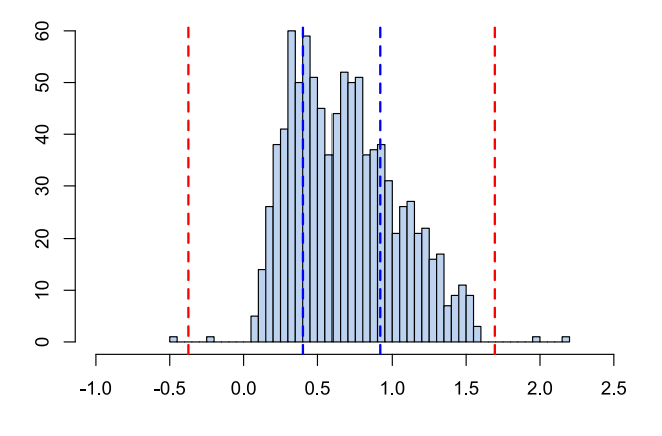
* Repetisjon fra PSY1010
* Bivariat regresjon
  + Linjer i planet
  + Målefeil
  + Tolkning av ustandardiserte og standardiserte koeffesienter
  + Minste kvadraters estimat
  + Kvadratsummer
  + Vurdering av underliggende antagelser
* Multippel regresjon (MR)
  + Partielle regresjonskoeffisienter
  + Statistisk kontroll
  + Supressoreffekter
  + Forklart varians i MR
  + Partiell korrelasjon
  + Polynomisk regresjon
  + Innflytelse
* Statistisk inferens i regresjon
  + Teststatistikker og samplingfordelinger
  + Standardfeil
  + t-fordelinger og F-fordelinger
  + P-verdier
    - Signifikans av individuelle prediktorer
    - Signifikans i forskjeller mellom modeller
  + Konfidensintervaller
* Avsluttende om regresjon
  + Kollinearitet
  + Kategoriske prediktorer
  + Interaksjon
  + Mediering
  + Modellbygging/Modellvalg
* En introduksjon til grunnleggende sannsynlighetsregning
  + Hva er utfallsrom, sannsynlighet, uavhengighet?
  + Fundamentale regler for sannsynlighetsregning (additiv, multiplikativ).
  + Kji-kvadrat statistikken og dens bruk i analyse av hvorvidt vi kan konkludere at to variabler er uavhengige.
* Variansanalyse
  + Regresjon med dummyvariabler
  + Sammenligning av gruppegjennomsnitt
  + Toveis variansanalyse
* Mer variansanalyse og bootstrapping
  + Litt mer om interaksjoner
  + Variansanalyse med repetert målinger
  + Bootstrapping

# Forelesning 1: Grunnleggende konsepter, og dataanalyse i R

**Varians**: mål på variabilitet rundt den forventede verdien (gjennomsnittet) på en fordeling.

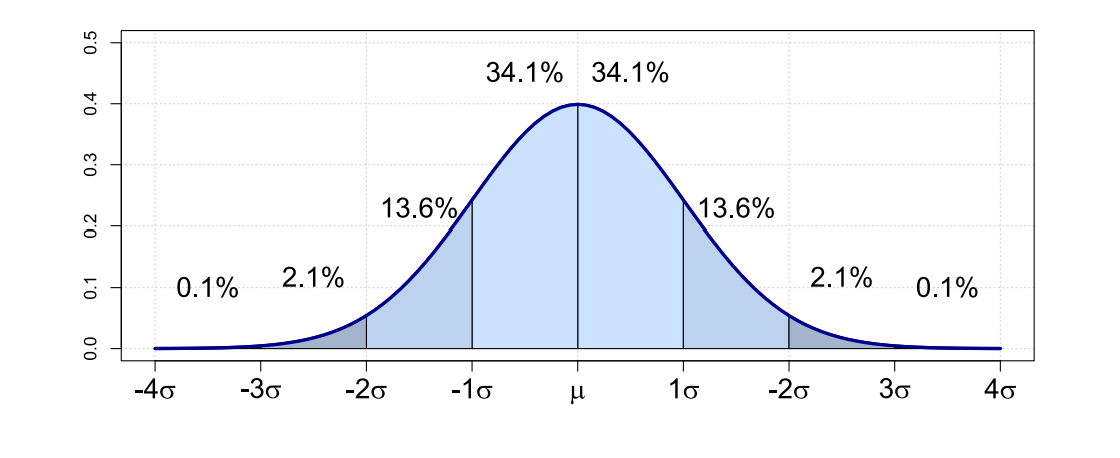


**Uteligger (outlier)**: En observasjon som ligger langt i verdi fra de øvrige observasjonene "Langt" er typisk definert som verdier større eller lik Q3+(1.5\*IQR), eller mindre enn Q1-(1.5\*IQR).



**Normalfordeling**: At en stokastisk variabel X er normalfordelt skriver vi ofte forkortet . Uttrykket for normalfordelingen er gitt ved: X ∼ N(μ, σ^2)

* Normalfordelings-antagelser:
  + Ca. 68% av fordelingen ligger innenfor ett standardavvik fra snittet.
  + Ca. 95% av fordelingen ligger innenfor to standardavvik fra snittet.
  + Ca. 99.7% av fordelingen ligger innenfor ett standardavvik fra snittet.

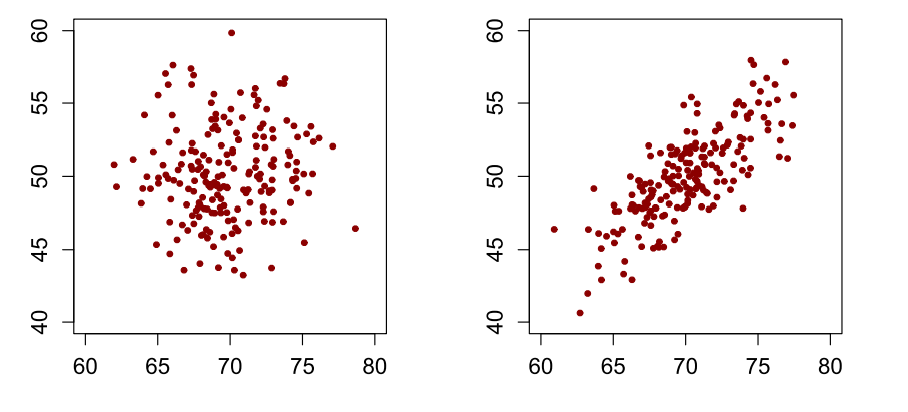


**Z-skåre**: Omregning av variabelen til enheter som utgjør antall standardavvik fra gjennomsnittet.

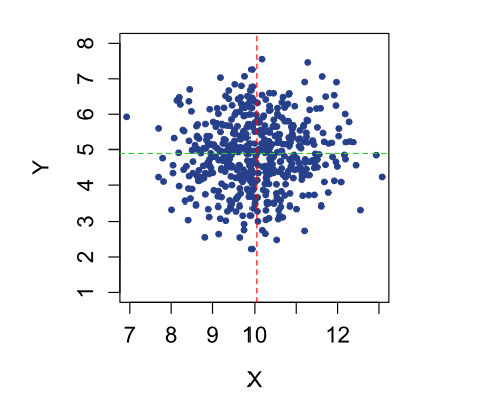
* Dersom en variabel , er X ∼ N(μ, σ^2), er Z ∼ N(0, 1)
  + dvs. en normalfordeling med snitt 0 og standardavvik 1.

**Spredningsplot:** En figur der hvert par med verdier på to variabler plottes som et punkt i planet.

* Uavhengig variable utgjør typisk X-verdien.
* Avhengig variabel utgjør typisk Y-verdien.

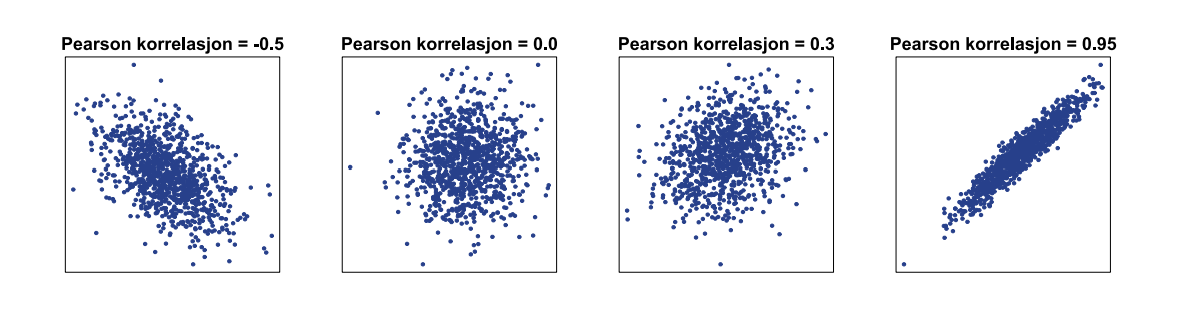


**Kovarians**: er en statistikk der grad av samvariasjon mellom to variabler kvantifiseres.



**Pearson-korrelasjon**:

* utgjør en standardisering av kovariansen.
* er definert som kovariansen delt på den største verdien denne kovariansen kan ta (produktet av variablenes standardavvik).
* er den vanligste statistikken brukt for å bedømme styrken av forholdet mellom to variabler.

****

**Forklart varians (r^2): Andelen varians to variabler har til felles**

* Et tall mellom 0 og 1 (siden korrelasjonen er mellom -1 og 1).
* Korrelasjon er derfor et mål på effektstørrelse (f.eks. om r = 0.5, så er 25% av variansen forklart)

# Forelesning 2: Bivariat regresjon

**Korrelasjon**

* Korrelasjon er lettere å tolke enn kovarians
* Kvadrering av korrelasjonen gir andel forklart varians

1. korrelasjoner er ikke robuste
   1. uteliggere kan ha en veldig stor effekt på korrelasjoner
2. korrelasjoner er påvirket av range
   1. Typisk synker korrelasjonen ettersom rangen til variablene minker.
3. Bakenforliggende variabler kan påvirke korrelasjonen
   1. Perason korrelasjoner er kun et mål på grad av «lineært» forhold mellom variablene.

* Kvantifiserer grad av lineær assosiasjon mellom to variabler.
* Tar verdier mellom -1 og 1. Er upåvirket av lineær transformasjon av variabler.
  + F.eks. deling og multiplisering
* Krever at begge variablene er kontinuerlige (ikke kategoriske).

**Rette linjer**

* En rett linje er unikt definert av to parametere:
  + : : konstantleddet / skjæringstallet (intercept)
    - Y-verdien når X er lik 0
  + : stigningstallet (slope)
    - Hvor mye endrer Y verdien seg når X øker med en enhet.

**Regresjonsfunksjon**: beskriver hvordan gjennomsnittet til en avhengig variabel følger verdiene til en uavhengig variabel.

**Bivariat regresjon**: vi har en avhengig variabel (Y) og en uavhengig variabel (X).

* Vi tenker at prosessen som ligger bak de observerte dataene er:
* Den forventede Y verdien gitt en X verdi basert på de estimerte regresjonskoeffisientene er gitt ved:
* er den "sanne" verdien i populasjonen, mens den estimerte verdien basert på et utvalg indikeres med en "hatt" .
* I uttrykket til den forventede verdien til en gitt person inngår ikke , fordi den forventede verdien er 0.

**Standardiserte regresjonskoeffisienter**:

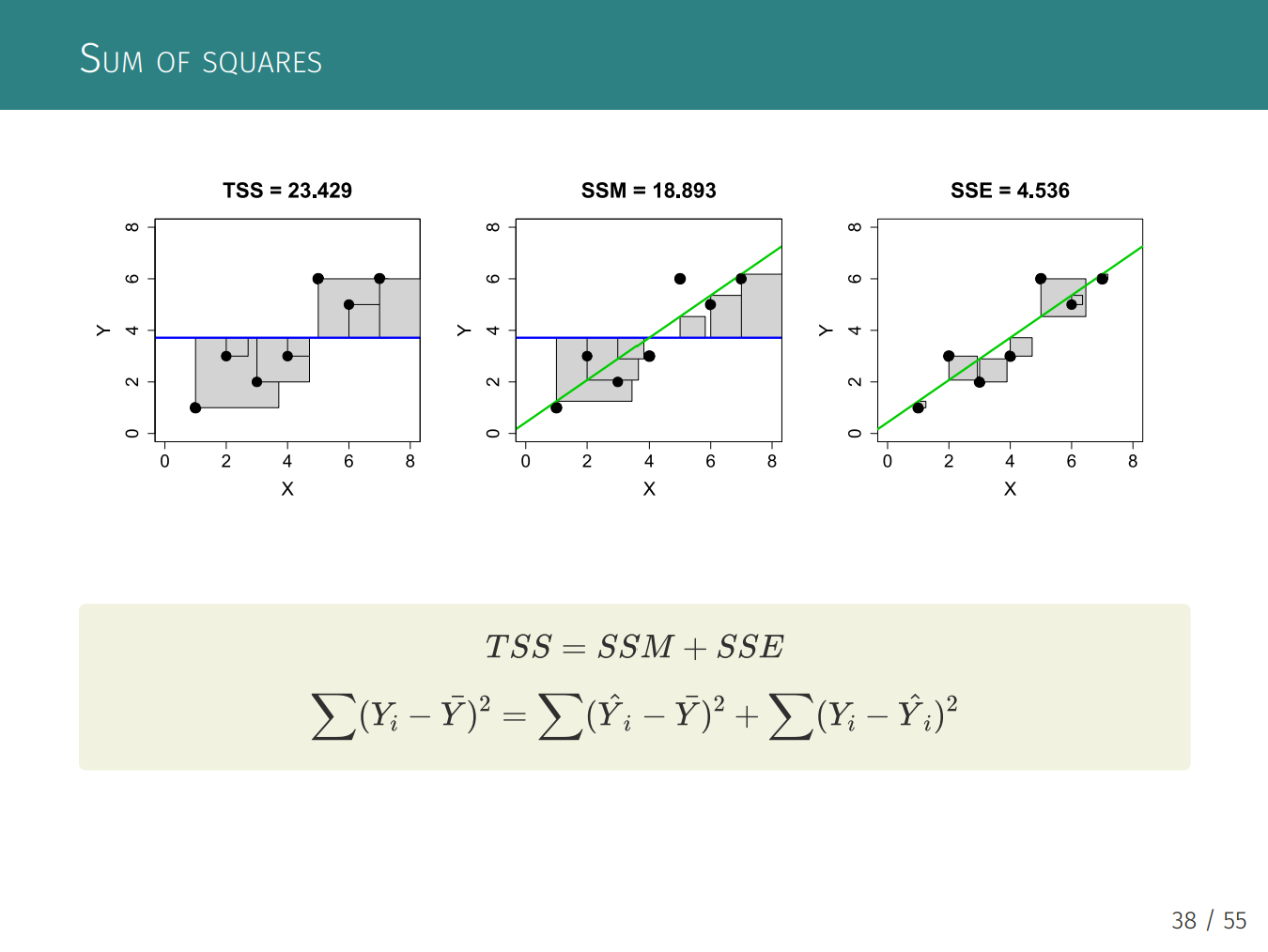
* Antall standardavvik endring i Y når X øker tilsvarende ett standardavvik.

**Residual**: Differansen mellom den observerte og den predikerte verdien.

**Minste kvadraters estimering**: Det minste kvadraters estimatet for b0 og b1 er verdiene b^hat0 og b^hat1 som minimerer summen av de kvadrerte residualene:

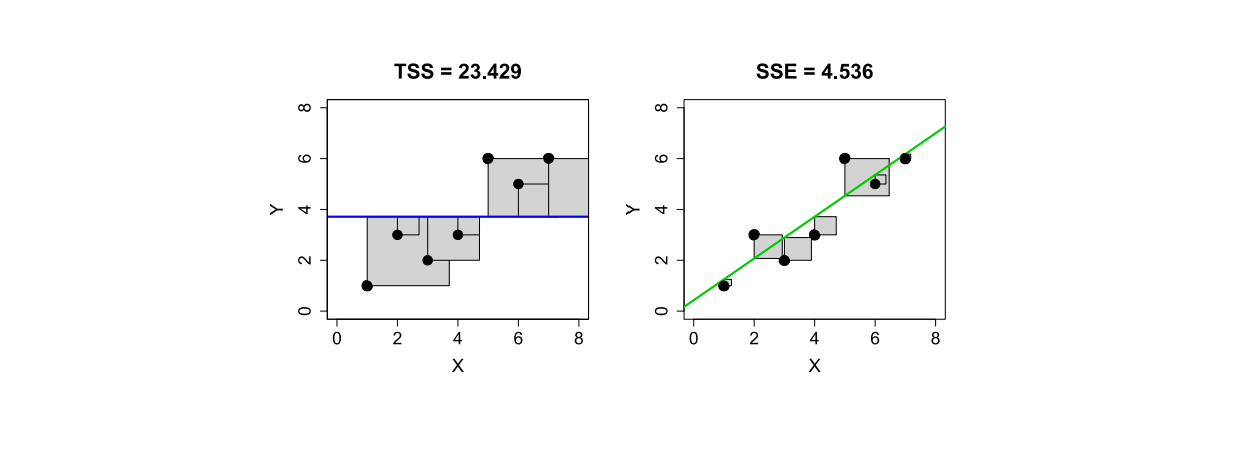
**Normal equations**: Det kan vises at minste kvadraters estimater av regresjonskoeffisientene kan regnes ut ved disse uttrykkene (kalles ofte normal equations).

**Sum of squares (TSS)**:

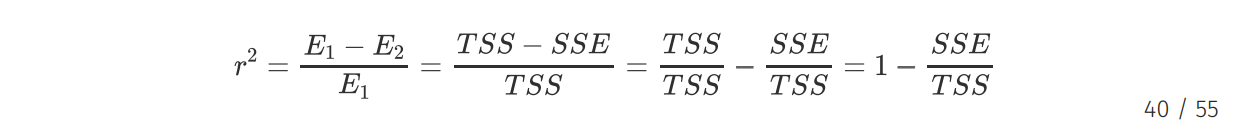


**Marginal varians**: Varians rundt gjennomsnittet. Dette er varians vi ikke kan forklare før vi tilpasser regresjonsmodellen.

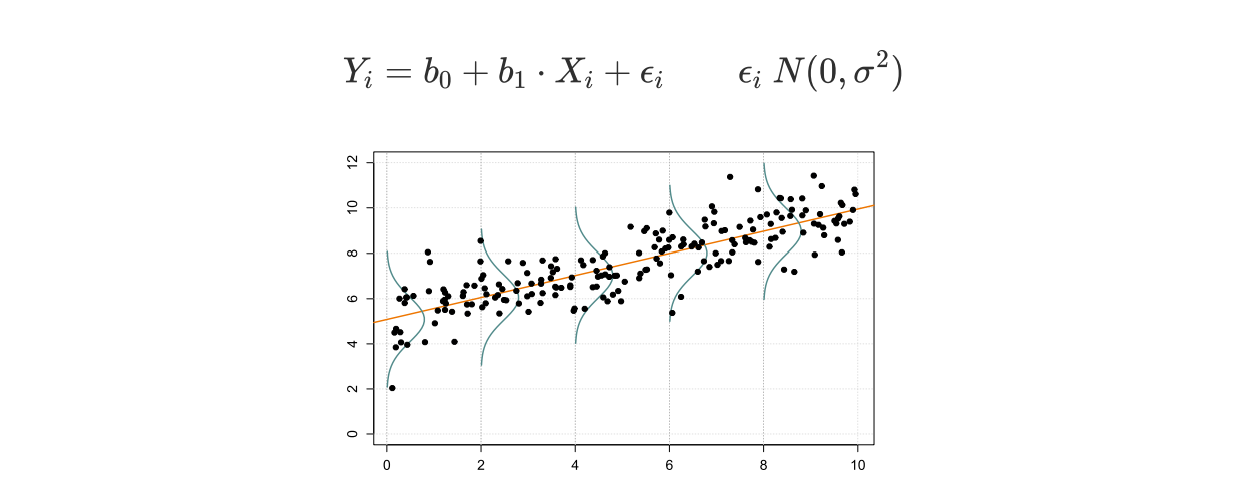
**Betinget varians**: Variasjon rundt regresjonslinjen. Dette er varians vi ikke kan gjøre rede for etter at vi har tilpasset regresjonsmodellen.

**Forklart varians (Coefficient of determination)**: 

* Kall summen av kvadrerte residualer i den naive modellen til venstre E1 , og summen av kvadrerte residualer i modellen til høyre E2 .
* r^2 er et mål på nedgangen i proporsjonen av uforklart varians når vi går fra en modell til en annen.



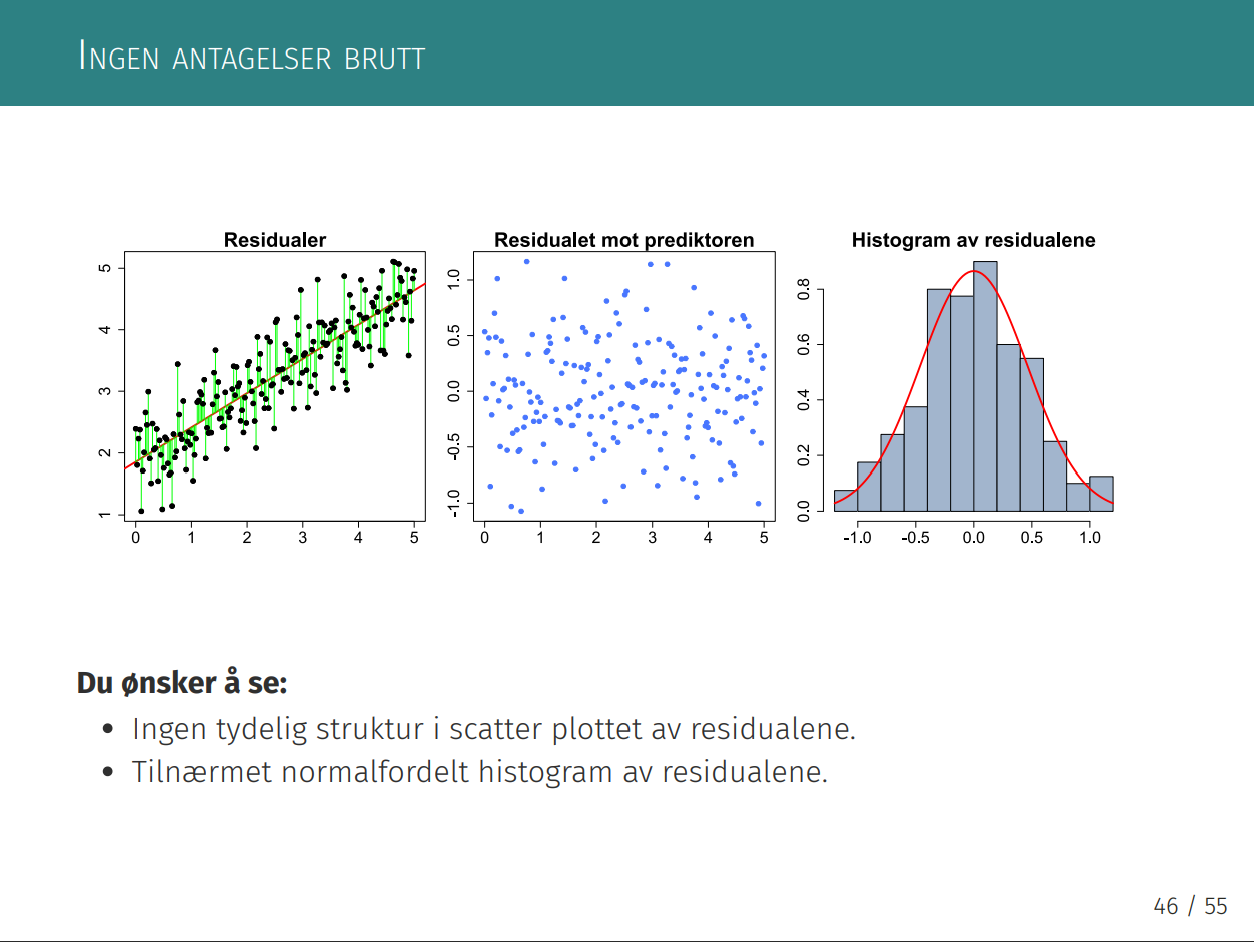
**Antagelser i regresjon**:

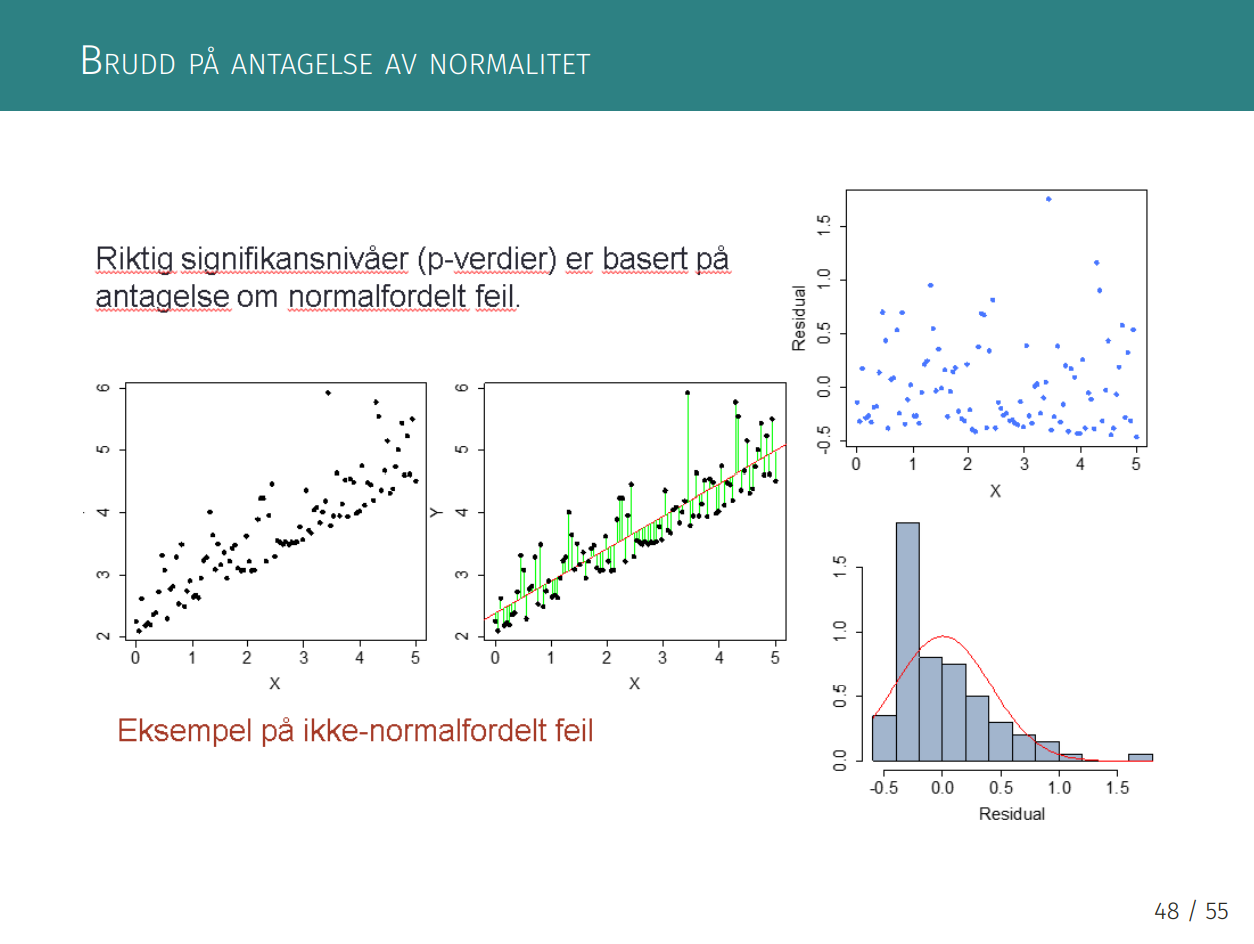
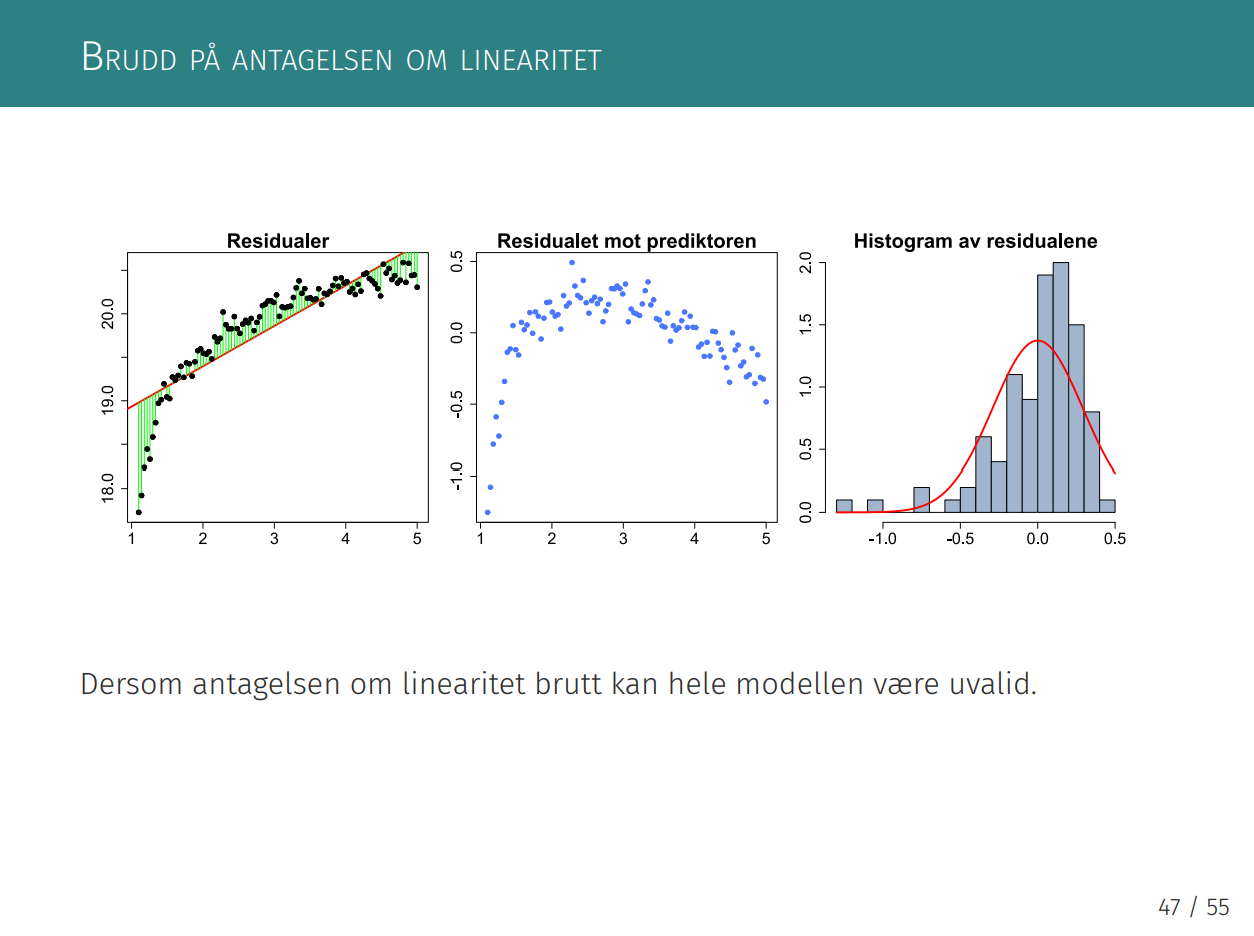


1. Forholdet mellom X og Y er lineært.
2. Feilvariansen (spredningen rundt regresjonslinjen) er normalfordelt, og har samme varians for alle nivåer av X.
3. Det er ingen ekstreme uteliggere.
4. Observasjonene er uavhengige.
5. Den uavhengige variabelen er målt uten feil

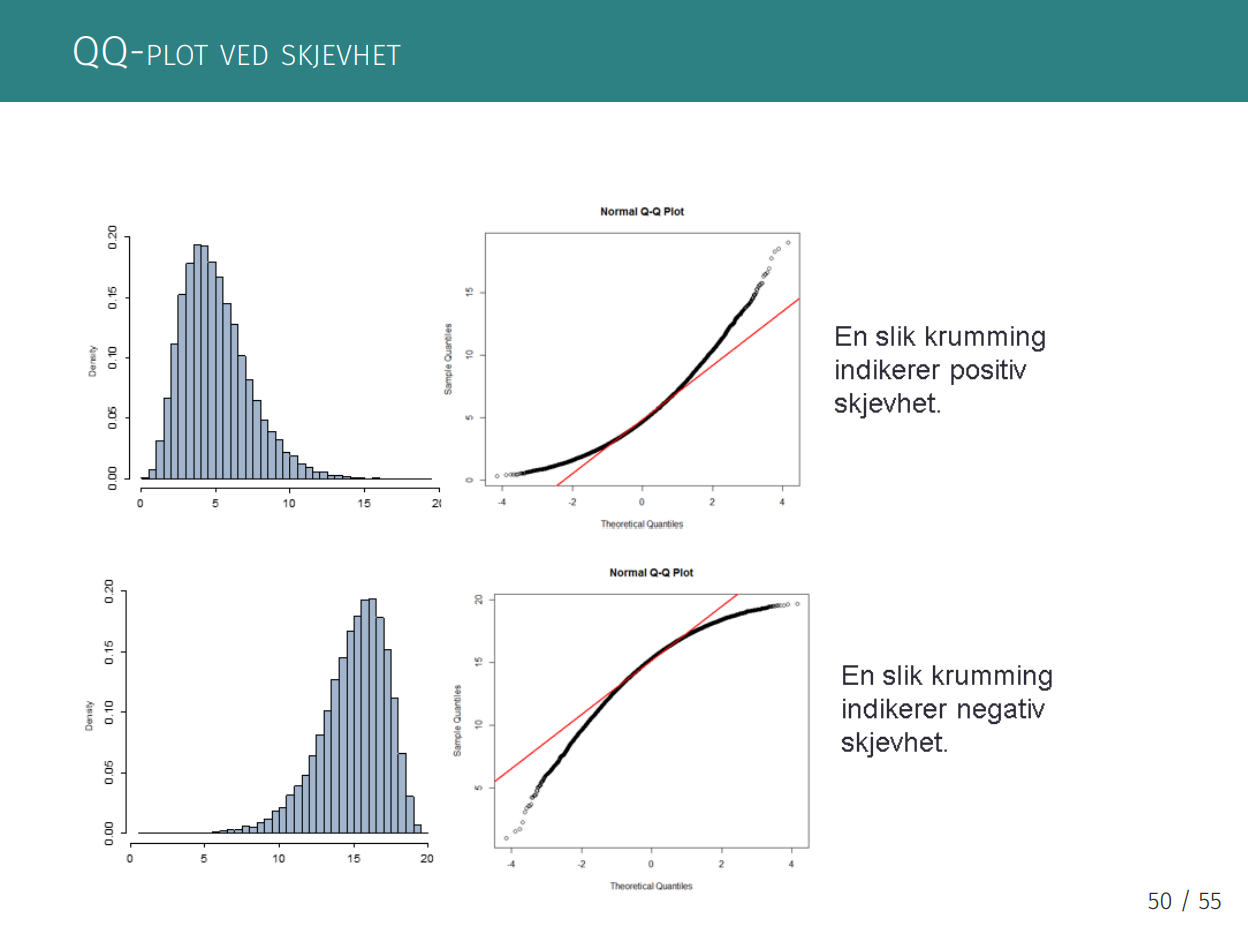
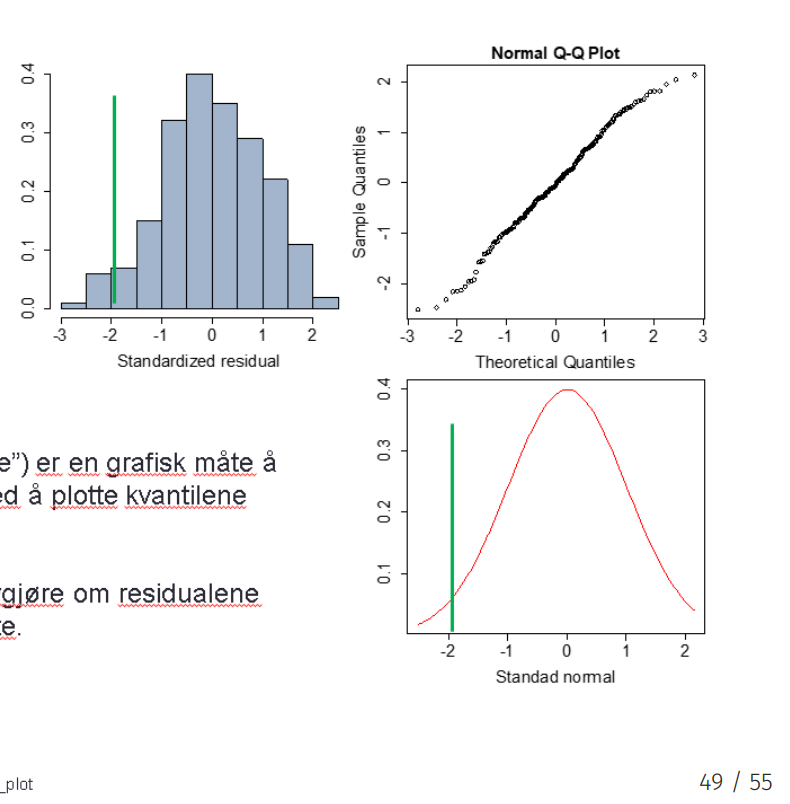
**Residual**: Differansen mellom den observerte og forventede verdien for en observasjon (Yi).

* Vi sjekker gjerne rimeligheten av de ulike antagelsene i lineær regresjon ved å inspisere diagnostiske plot av residualene



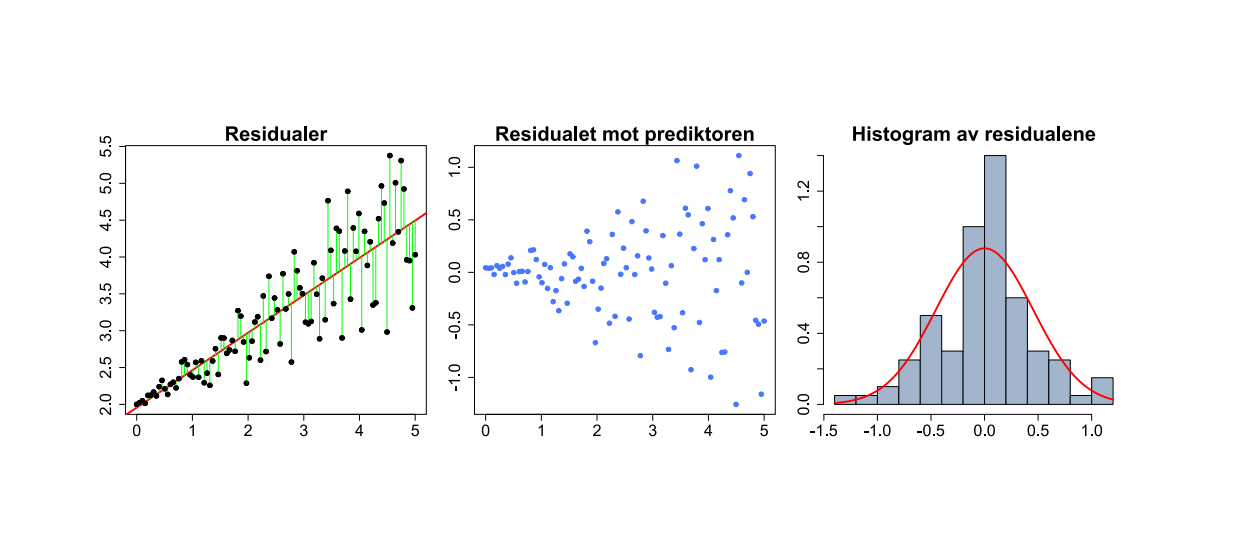


**QQ-plot**: grafisk måte å sammenlikne to fordelinger ved å plotte quantilene deres mot hverandre. Vi vil typisk bruke den for å avgjøre om residualene kan sies å være normalfordelte.



**Brudd på antagelsen om feilvarians**:

* Er antagelsen om konstant feilvarians brutt vil du få valide koeffisienter, men trolig gale p-verdier (og konfidensintervaller).



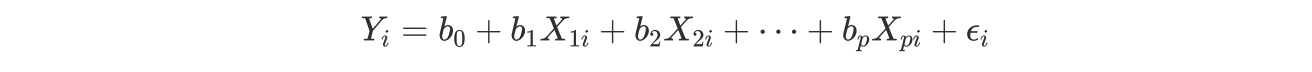
# Forelesning 3: Multippel regresjon

**Hvorfor ha mer enn én uavhengig variabel?**

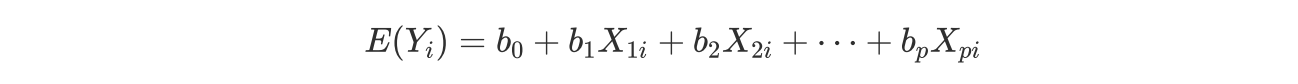
1. Mer nøyaktig prediksjoner
   1. Den avhengige variabelen er hovedfokus.
   2. F.eks. Predikere bedre hvem som dropper ut av skolen.
2. Isolere effekten av en uavhengig variabel (statistisk kontroll)
   1. Interessen er i en eller flere spesifikke uavhengige variabler.
   2. F.eks. Øker fattigdom risikoen for frafall?
3. Forstå den samlede effekten av flere uavhengige variabler
   1. Flere uavhengige variabler kan samlet påvirke utfallet på kompliserte måter.
   2. F.eks. Er effekten av fattigdom på frafall sterkere for barn av enslige foreldre (interaksjon).
   3. F.eks. Fører fattigdom til frafall fordi foreldre i fattige familier har mindre tid til å følge med på barna (mediering).

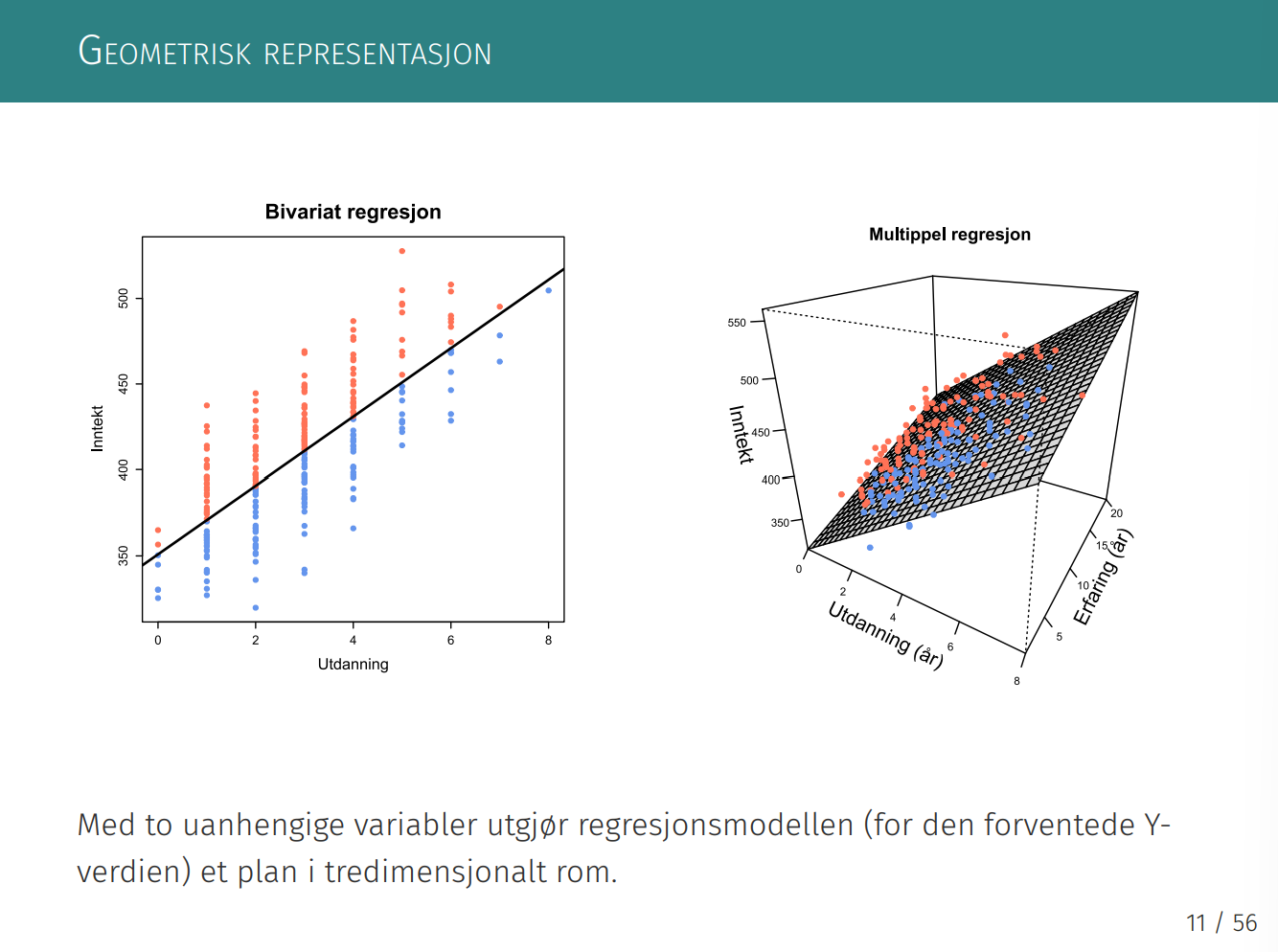
**Multippel regresjon**:

* Har vi to uavhengige variabler X1 og X2 , refererer vi til person i’s skårer på disse variablene som X1i og X2i (f.eks.utdanning og arbeidserfaring)
* Med p uavhengige variabler blir regresjonsuttrykket:

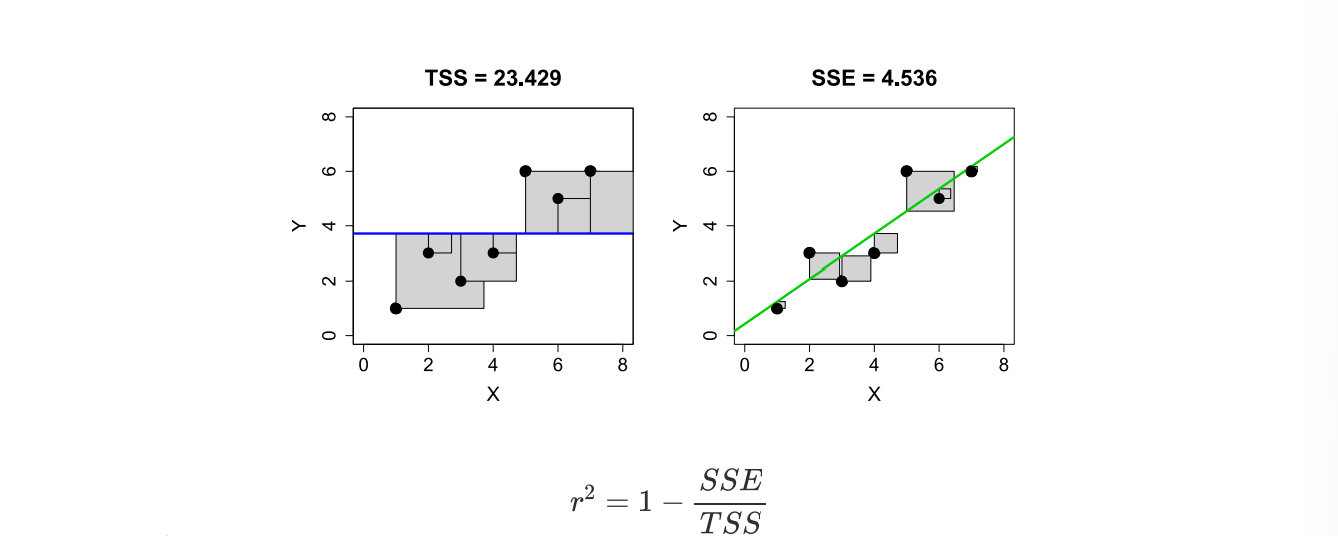


* Parametrene (bj ) kalles partielle regresjonskoeffisienter.
  + De tolkes som forventet endring i Y som følger av at Xj endres en enhet, men alle andre uavhengige variabler holdes konstant.
* Den forventede verdien for Y for individ i basert på verdiene på de uavhengige variablene er gitt ved:



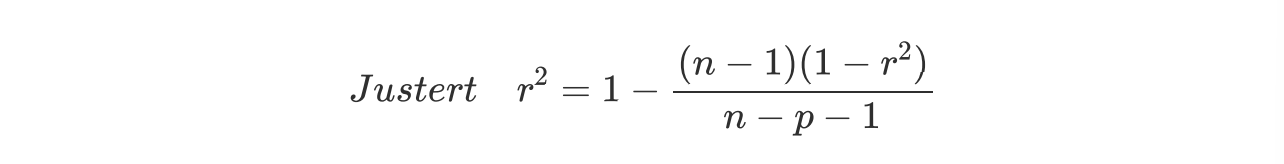


**Forklart varians (Coefficient of determination)**: r^2 er et mål for nedgangen i proporsjonen av uforklart varians når vi går fra en modell til en annen.



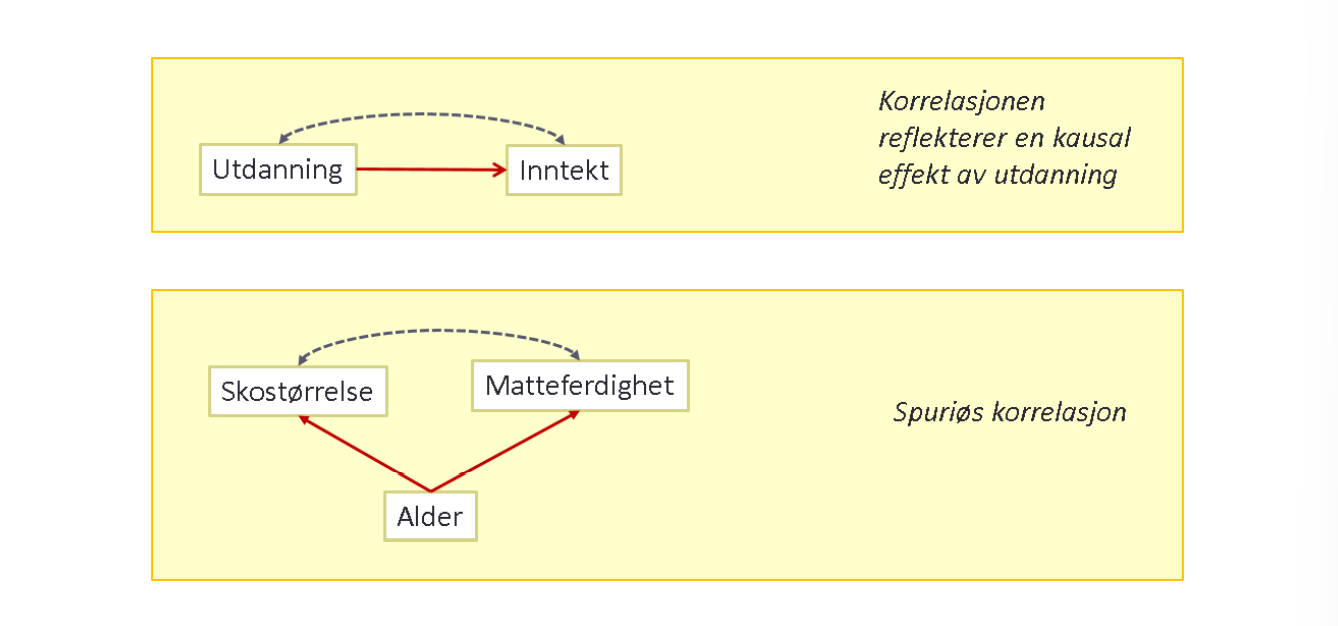
**Justert r^2**: en statistikk som korrigerer r^2 for utvalgsstørrelse og antall uavhengige variabler i modellen.

* Når n (antall observasjoner) er stor og p (antall uavhengige variabler) er liten, så er r^2 og justert r^2 praktisk talt identiske.



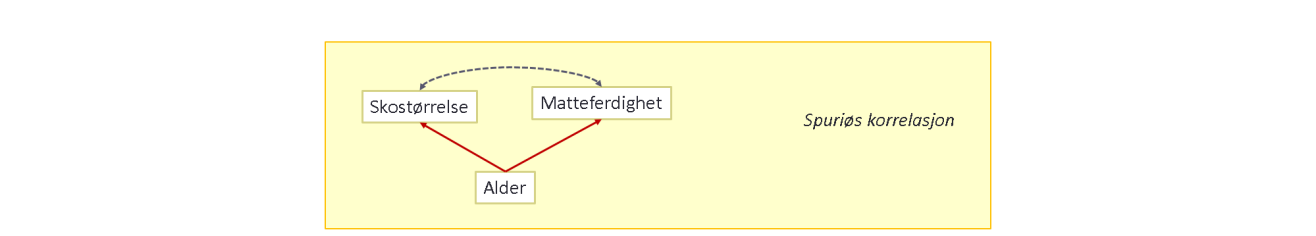
* r^2 øker alltid om du legger til flere uavhengige variabler, men justert r 2 kan synke.
* Når de uavhengige variablene er ukorrelerte forklarer de «separate proporsjoner» av variansen i den avhengige, og forklart varians i en modell med begge er summen av hva de kan forklare enkeltvis.
* Når to uavhengige variabler er korrelerte, vil den variansen de kan forklare samlet være mindre enn summen av hva de kan forklare enkeltvis.

**Problemet med tredjevariabler**:



* Utenforliggende tredjevariabler kan skape en spuriøs korrelasjon.

**Spuriøs**: Falsk, brukt hovedsakelig innen epidemiologi for å beskrive en relasjon som virker kausal, men som ikke er det.

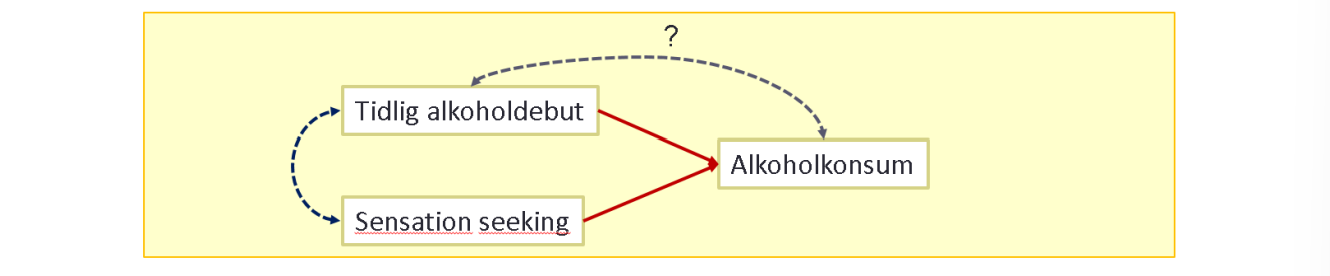


**Utenforliggende variabel (extraneous)**: andre variabler enn den uavhengige variabelen i et eksperiment som påvirker (eller er assosiert med) den avhengige variabelen.

* F.eks. for inntekt; kjønn, personlighet, ...
* Dette er variabler som (du kanskje ikke har målt), men som står for noe av den uforklarte «erroren» rundt regresjonslinjen.

**konfunderende variabel**: er en utenforliggende variabel som er assosiert med (korrelerer med) både den avhengige og den uavhengige variabelen du er interessert i.

* Dette skaper kan skape en tilsynelatende kausal sammenheng mellom eksponeringen og utfallet.



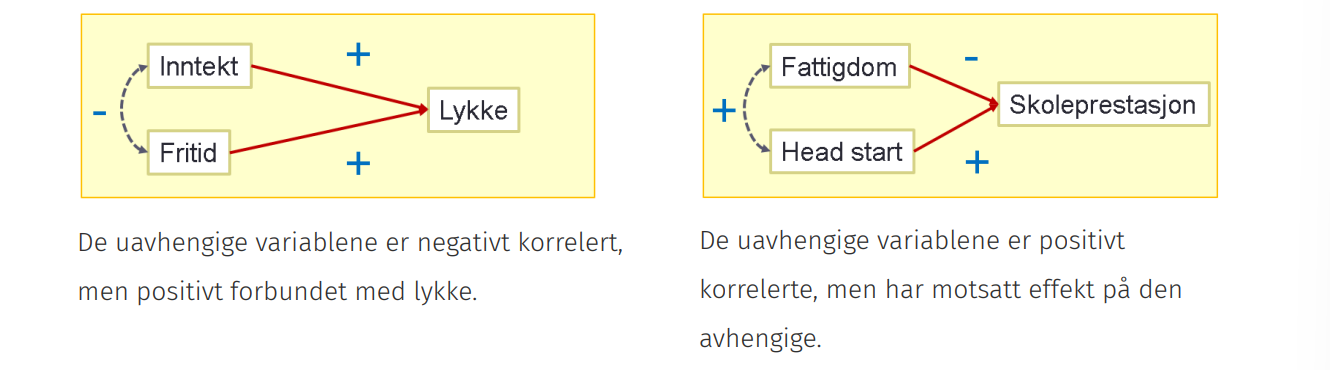
* Konfunderende variabler korrelerer altså med både eksponeringen og utfallet.
* Er du interessert i sammenhengen mellom tidlig alkoholdebut og senere alkoholkonsum er sensation seeking er en mulig konfunderende variabel.

**Statistisk kontroll**:

* I eksperimentelle studier fjerner vi effekten av konfunderende variabler ved å holde dem konstante.
  + F.eks. vi inkluder bare deltagere med samme sensation seeking skåre.
  + Har alle samme skåre på sensation seeking, kan dette ikke forklare hvorfor folk varierer i alkoholkonsum.
* I observasjonelle studier trenger vi også en måte å holde en «konfunderende variabel konstant».
  + Dårlig løsning: Bare se på effekten innad i et subsett av deltagere som har samme skåre på sensation seeking.
    - Dårlig, fordi vi kaster bort det meste av dataene.
  + Bedre løsning: Inkluder både sensation seeking og alder for alkoholdebut som uavhengige variabler i regresjonsmodellen.
    - For å se hvorfor dette hjelper oss å holde sensation seeking konstant hjelper det å ha en geometrisk modell for regresjonsplanet.

**Suppressorer**: Noen ganger er det lite/ingen assosiasjon mellom to variabler inntil du kontrollerer for en tredje.

* Den tredje variabelen kalles da en supressor. Den «suppress» (fortrenger) assosiasjonen mellom de første to.



**Standardiserte regresjonskoeffisienter**:

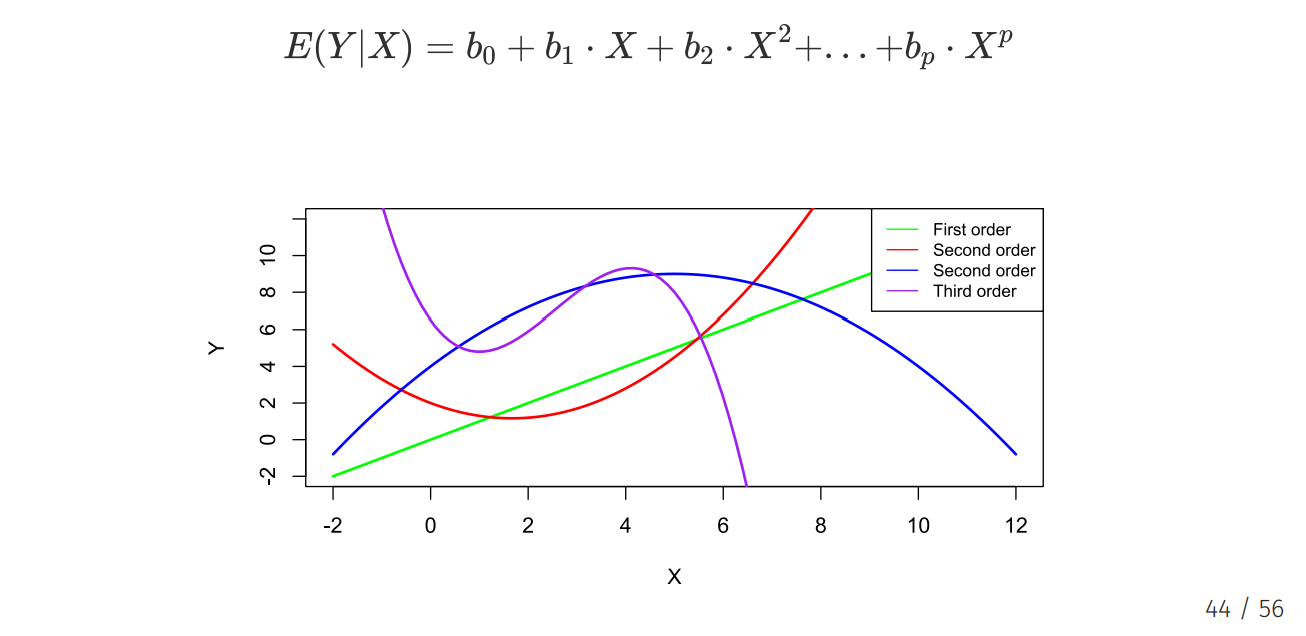
* Det gir som regel ikke mening å sammenlikne de ustandariserte koeffisientene, fordi variablene kan ha veldig ulik skala.
* For å kunne sammenlikne regresjonskoeffisienter må vi sette dem på en liknende skala, der "en enhet økning" er sammenliknbare størrelser.

**Partiell korrelasjon**: korrelasjonen mellom to variabler der den effekten som overlapper med en eller flere andre variabler er fjernet.

* , korrelasjonen mellom y og x1 kontrollert for x2 .

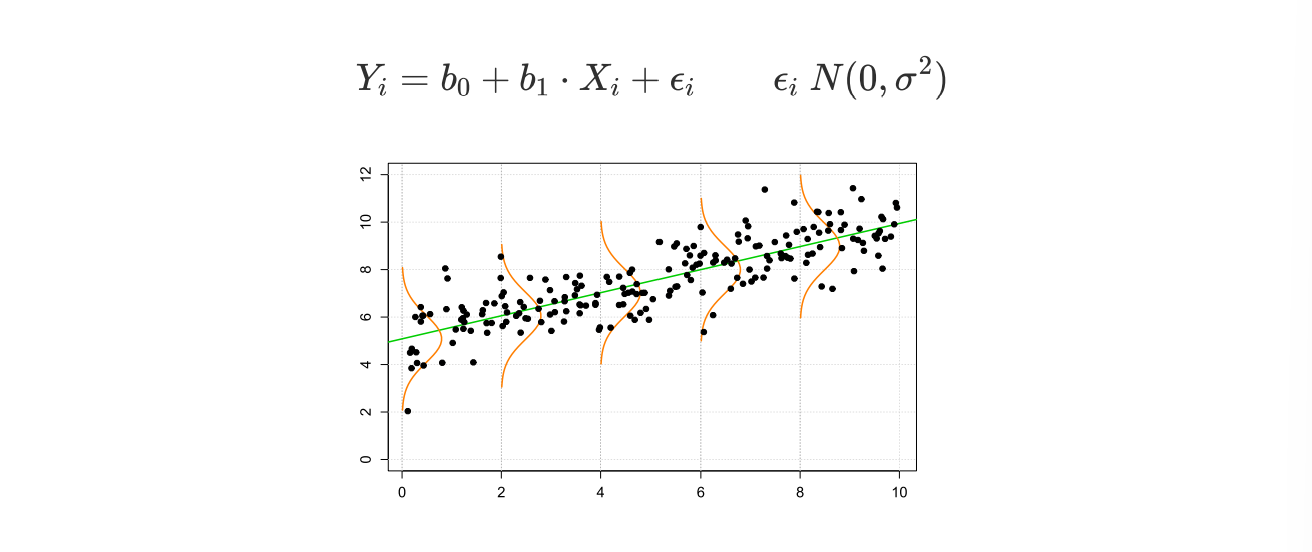
**Ikke-lineære forhold og polynomisk regresjon**:

* 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 en polynomisk regresjonsfunksjon.



* Obs ved polynomiske modeller
  + Vær kritisk til modeller med høy polynomer, de kan utgjøre en overtilpasning til dataene.
  + Modellen kan ha begrenset gyldighet utenfor ranger der vi har observerte verdier av den uavhengige variabelen (unngå ekstrapolering).
  + Koeffisientene har ikke lenger tolkningen partielle stigningsgrader. Det er ikke mulig å se på alder, men "holde alder^2 konstant".

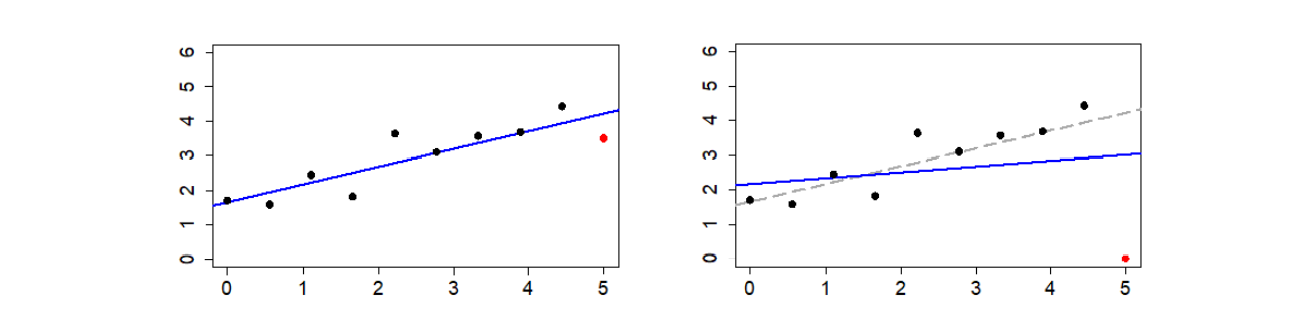
**Antagelser i regresjon**:



1. Forholdet mellom X og Y er lineært.
2. Feilvariansen (spredningen rundt regresjonslinjen) 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 målt uten feil.

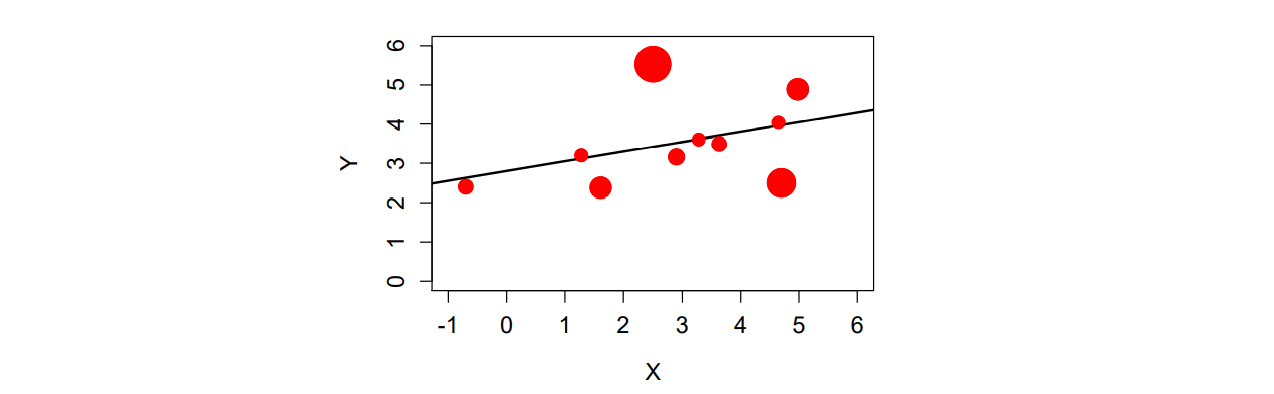
**Uteligger**: en observasjon som ligger lang i verdi fra de øvrige. uavhengige variabler.

* + F.eks. Det er ikke uvanlig å være 150cm eller veie 120 kg., men individer som både er 150cm høy og veier 110 kg er sjeldne.
  + Uteliggere blir identifisert ved ulike distansemål.
* Det vanligste målet på avstand fra modellen er residualet, . Dette er nyttig for å identifisere uteliggere på den avhengige variabelen.



**Residualer**:

* Det vanligste målet på distanse (avvik) er residualet, . Størrelsen på residualene hjelper oss å avdekke uteliggere på den avhengige variabelen



**Influens**: Mål på influens kombinerer avstand og leverage for å identifisere særlig innflytelsesrike observasjoner.

* Det vanligste målet på innflytelse er Cook’s D, som kvantifiserer grad av endring du ville sett i stigningstallet til linjen dersom en observasjon var slettet. Observasjoner med verdier > 1 bør granskes.
* Andre vanlige influens statistikker; DFFIT and DFBETA.

# Forelesning 4: Inferens i regresjon

**Slutningsstatistikk**: Det å trekke slutninger om populasjoner gjennom egenskaper ved utvalg.

* F.eks. dersom gjennomsnittlig IQ i et utvalg er 110, kan du være helt sikker på at utvalget er trukket fra en populasjon der snittet er over 100?

**t-statistikken**: som brukes til å teste om individuelle regresjonskoeffisienter et statistisk signifikant.

**F-statistikken**: brukes til å teste den samlede innflytelsen til flere uavhengige variabler.

**Samplingfordelingen**: til en statistikk er fordelingen av de verdiene statistikken tar på tvers av ulike utvalg (av lik størrelse, og trukket fra samme populasjon).

* Standardavviket til en samplingfordeling kalles standardfeilen (SE).
* Fordelingen over kalles også gjerne nullfordelingen
  + Viser den relative sannsynligheten til ulike antall kron når nullhypotesen er sann (altså her at mynten er rettferdig).

**Sentralgrenseteoremet (central limit theorem)**: summen av avhengige og identisk fordelte variabler er nærmer seg en normalfordelt ettersom utvalgsstørrelsen vokser.

* Dette er det viktigste teoremet innen statistikk, og basisen for nesten all klassisk slutningsstatistikk.
* Moralen er: Noen ganger vet vi (gjennom sentralgrenseteoremet) hva samplingfordelingen til vår statistikk er.

**p-verdien**: sannsynligheten for å en verdi på teststatistikken som er lik elle større enn den observerte (gitt at H0 er sann).

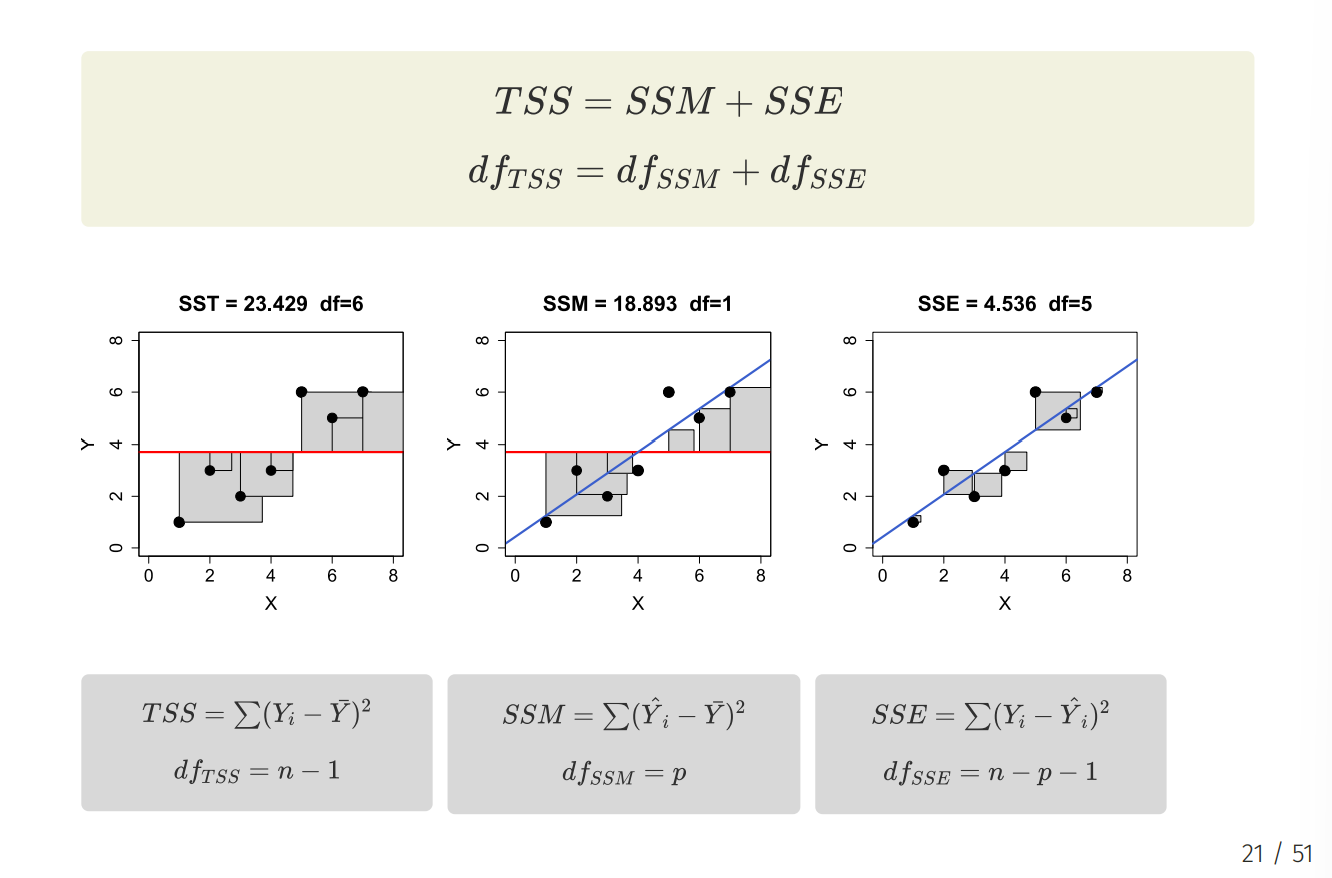
**Frihetsgrader**: Frihetsgrader (df) til en statistikk antall verdier i statistikken “får lov til” å variere.

* 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.

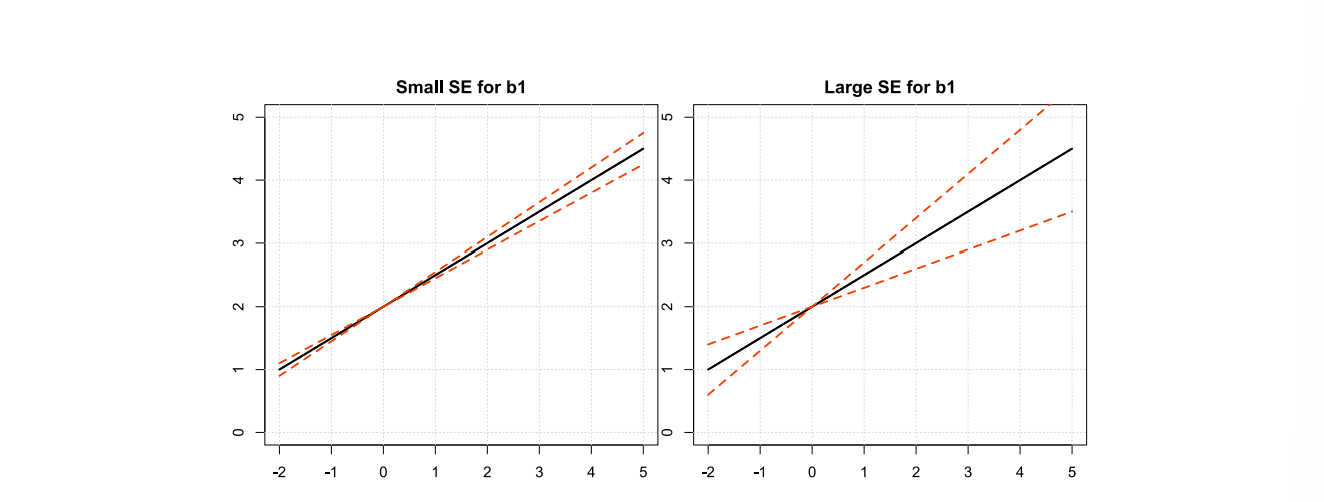
Et bilde som inneholder tekst

Automatisk generert beskrivelse

**Kvadratsummer med frihetsgrader (p: antall uavhengige variabler)**

****

**Standardfeilen til b1**

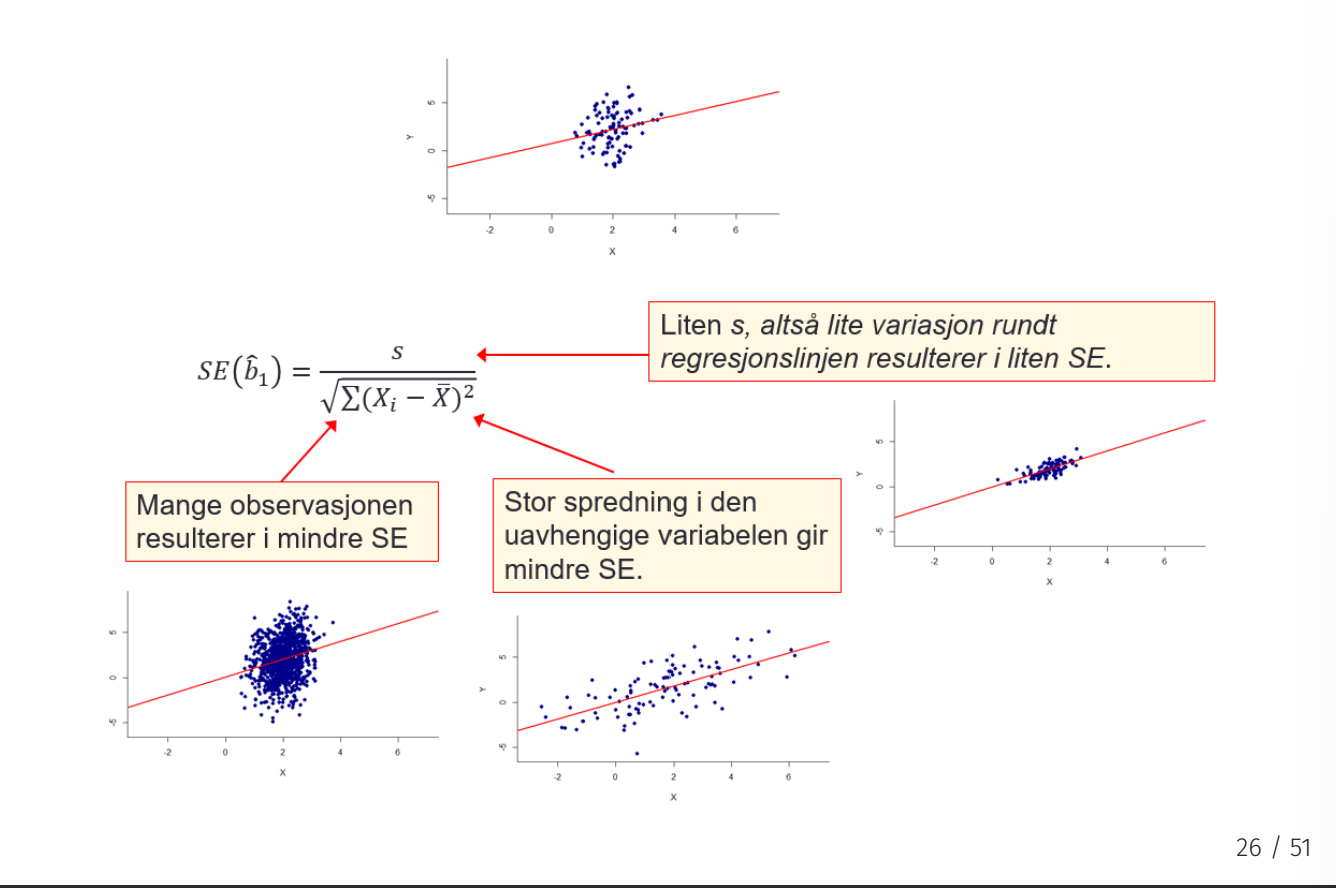


* (SE(\hat{b}\_1)) er et estimat på variabiliteten (standardavviket) i stigningstallet du ville ha sett på tvers av ulike utvalg tatt fra den samme populasjonen.
  + Du kan bruke ditt utvalg til å estimere denne variabiliteten

**Residual standard error (conditional standard error)**: et estimat av i uttrykkeEt bilde som inneholder tekst

Automatisk generert beskrivelse

**standard error of the slope**:



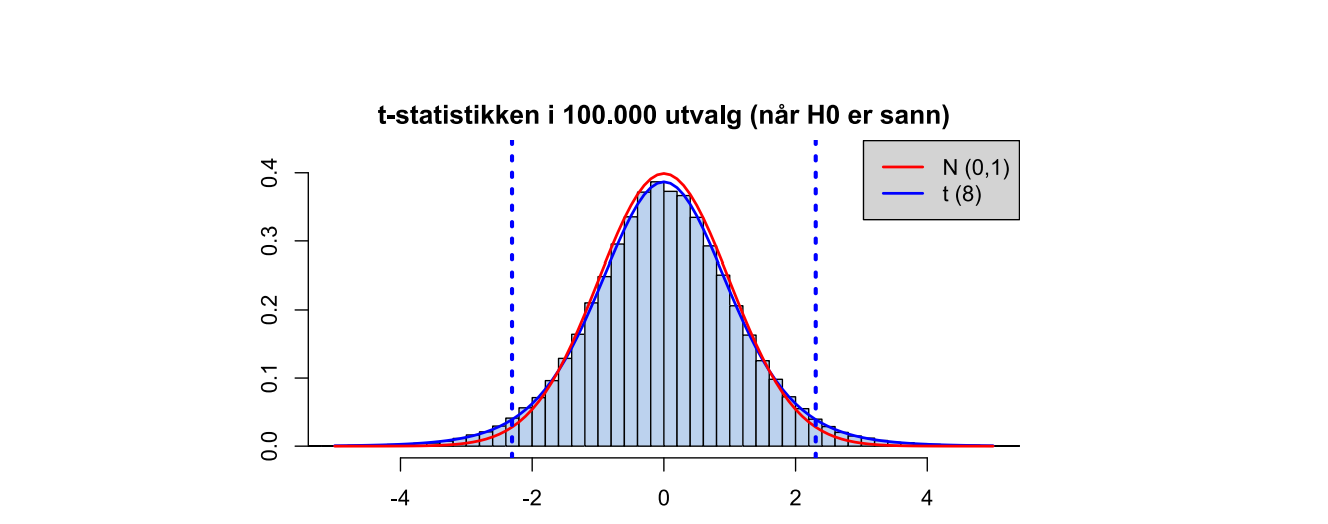
**t-statistikken**:

* Samplingfordeling under H0: t(n-p-1)

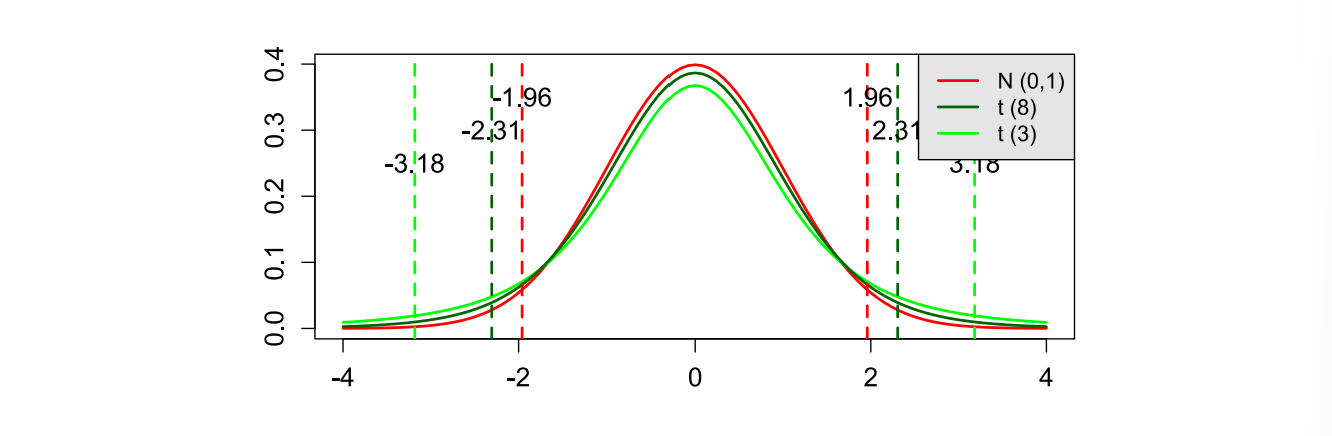
**t-fordelingen**:

* 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, nærmer en t-fordeling seg en standard normalfordeling, og de er praktisk talt identiske når df > 30.

**Kritiske verdier**:

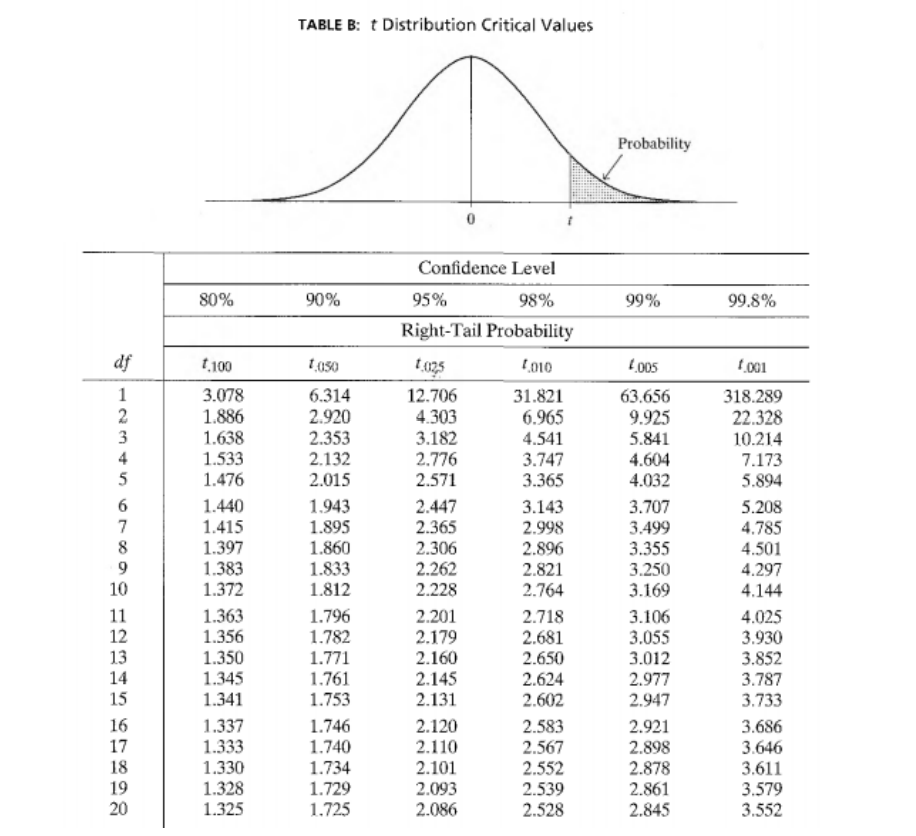


* Når nullhypotesen er sann (b1=0), så følger t-statistikken i ulike tilfeldige utvalg på størrelse 10 en t-fordeling med 8 frihetsgrader.
* Jeg kan da bruke en tabell for å finne de kritiske verdiene til t-statistikken.
* Under H0 er t-statistikken t-fordelt med (n-p-1) frihetsgrader.



* While in a standard normal distribution, 95% of the area is between +-1.96, on a t(8) distribution, 95% is between +-2.31 and for a t(3) distribution, +- 3.18.
* Moral: with fewer degrees of freedom there is more uncertainty.

**t-tabellen**:

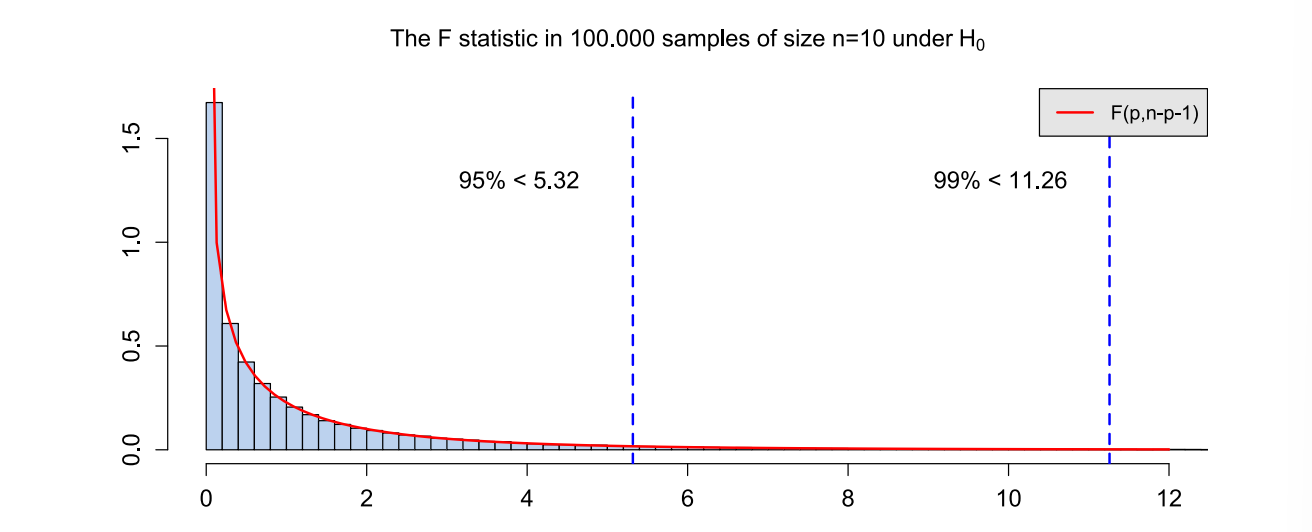


**Konfidensintervaller**:

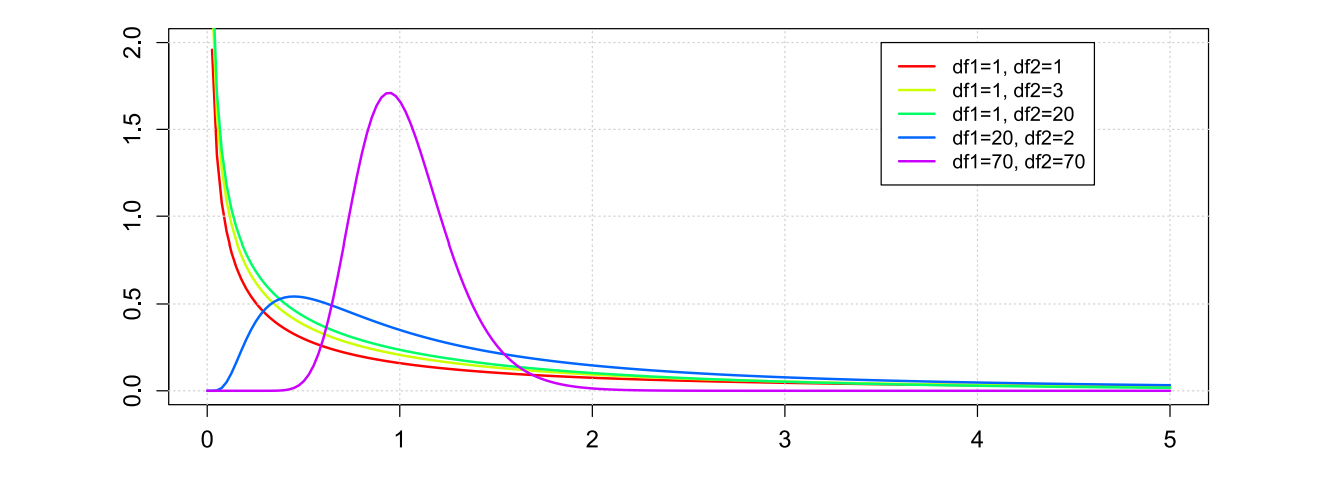
* Konfidensintervaller (CIs) er en annen måte å uttrykke usikkerhet ved estimatene.
* Enhver verdi innad i et 95% konfidensintervall utgjør en mulig hypotese om parameterets Verdi som du ikke kan forkaste på et 0.05 nivå.
* Enhver verdi utenfor intervallet kan forkastes på et 0.05 nivå.
* Konfidensintervall gir mer informasjon enn en p-Verdi.
  + Om 0 er utenfor intervallet betyr det at vi kan forkaste hypotesen om at parameteret har verdien 0.
  + I tillegg kan vi se hvilke andre verdier vi kan forkaste.

**F-testen**: Testing av flere regresjonskoeffsienter samlet.

* Samplingfordeling under H0: F(p,n-p-1)



* For the F-distribution, only a one tailed test is reasonable.



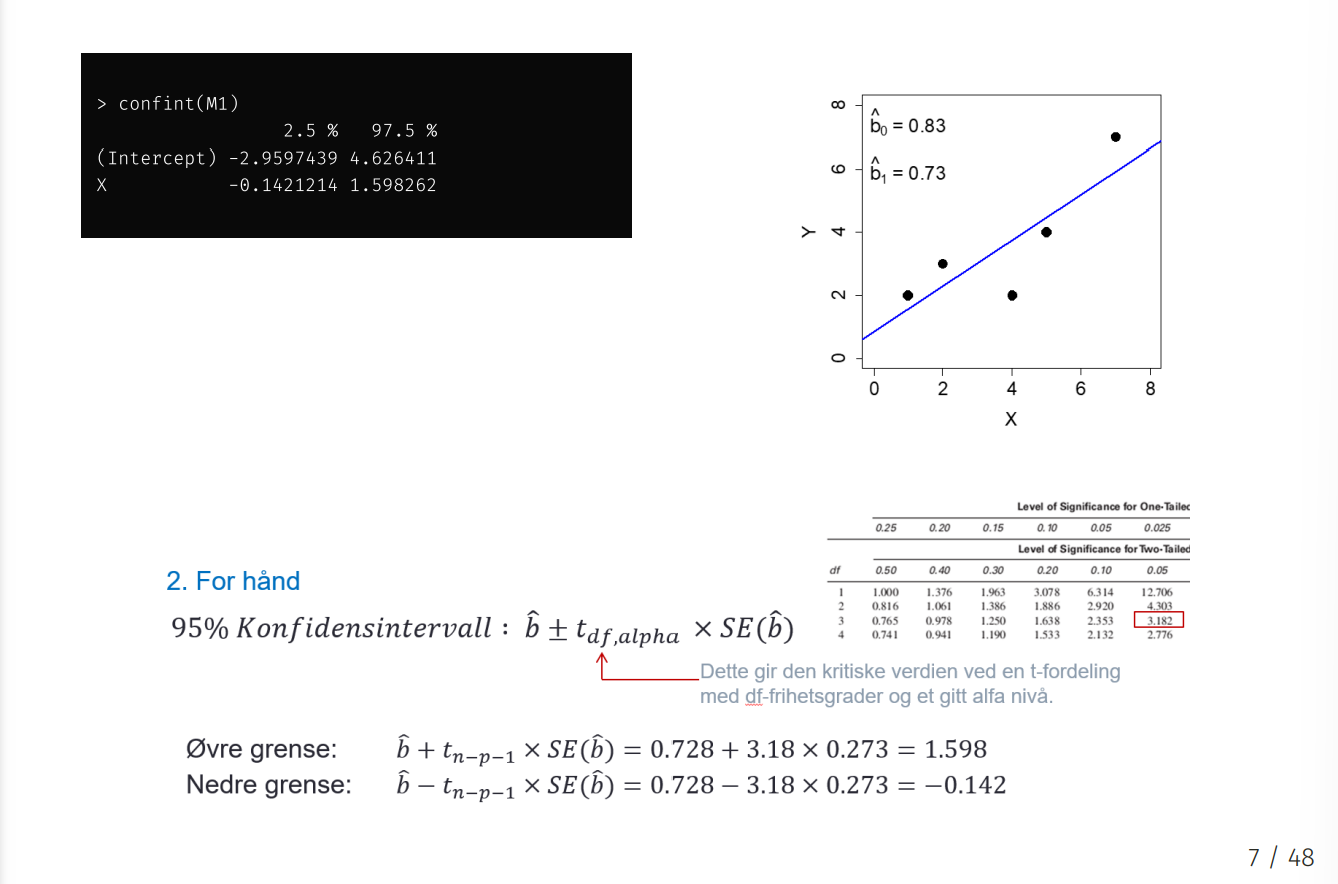
**F-tabellen**:

Et bilde som inneholder bord

Automatisk generert beskrivelse

# Forelesning 5: Kategoriske prediktorer, interaksjon, mediering

**Utregning av konfidensintervaller**:



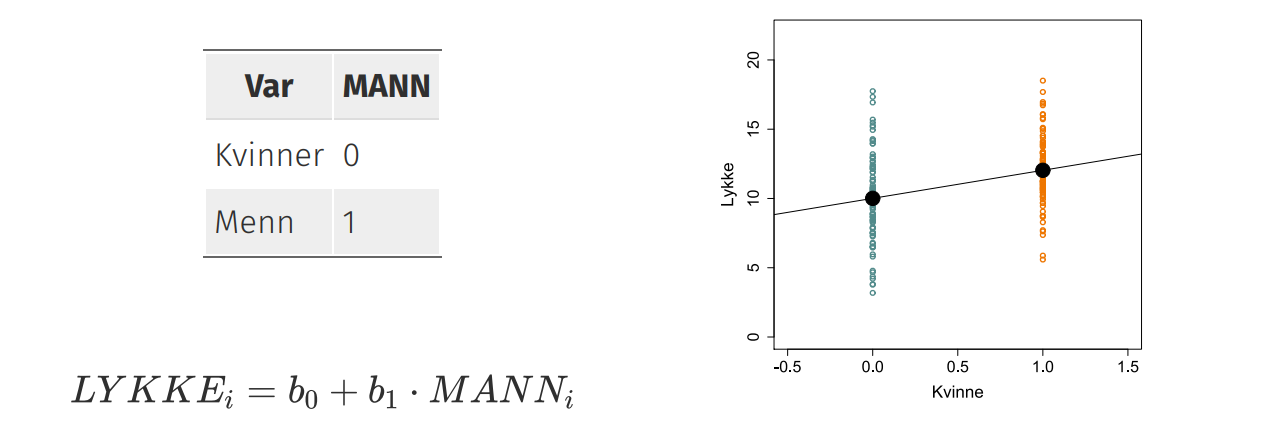
**Kollinearitet**: graden av lineær sammenheng mellom flere forklaringsvariabler i en multippel regresjonsmodell. Blir denne for høy (korrelasjon >>0.8) kan det medføre problemer.

* Konsekvenser av kollinearitet
  + 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.

**VIF (variance inflation factor)**: diagnostisk statistikk som angir grad av multikollinearitet.

* VIF tallfester hvor mye en uavhengig variabel øker standardfeilen
* Høye VIF-verdier (over 4 eller 5) tyder på kollinearitetsproblemer

**Dummyvariabel** (indikatorvariabel): tar bare verdiene 0 eller 1.



**Kategorisk 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 regresjonsmodeller kan vi rekode dem til dummy-variabler.
* I kodingen under blir Oslo referansekategori.

**additive effekter**: Effekter er additive dersom den kombinerte effekten av et lik summen av enkelteffektene.

* Universitetsgrad gir 1000K NOK forventet økt lønn og 3 år arbeidserfaring gir 50K NOK. Effekten er additiv om forventet lønn for en med både universitetsgrad og 3 år arbeidserfaring er 150K NOK.
* Genvariant X bidrar til 0.5 cm høyde, genvariant Y bidrar til 0.3 cm høyde, effekten av å ha begge variantene er 0.8 cm høyde.

**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.

* Genvariant X eller Y resulterer ikke I sykdom alene. Sykdom oppstår kun når du arver begge variantene.

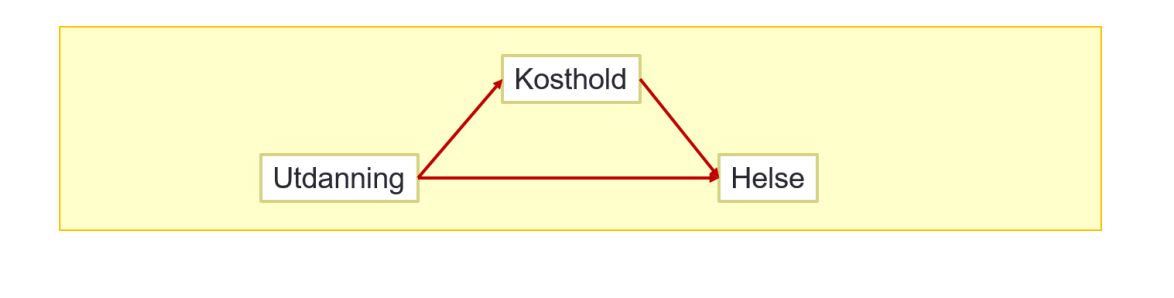
Et bilde som inneholder tekst, klokke

Automatisk generert beskrivelse

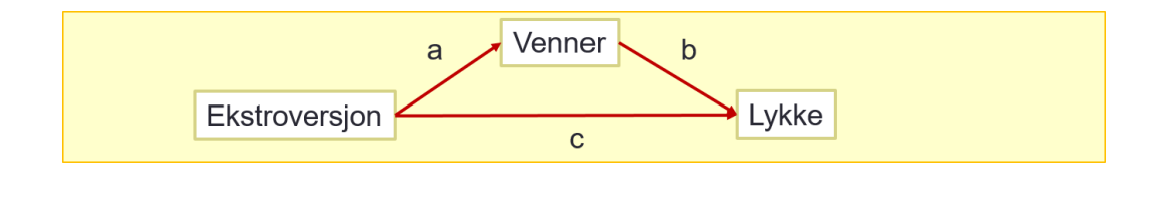
**Kryssproduktledd**: For å modellere interaksjon vil jeg definere et kryssproduktledd som er angitt av produktet mellom verdiene på de to uavhengige variablene.

**Mediator**: er en variabel som formidler deler av effekten/risikoen av andre uavhengige variabler på utfallet.

* På rent statistik 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.



**Variabel-assosiasjoner**:

Et bilde som inneholder bord

Automatisk generert beskrivelse

**Modellvalg**: Hvorvidt en uavhengige 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 som skal inngå i den endelige modellen?

1. Hierarkisk regresjon
   1. Samle uavhenige variabler i grupper (blokker), og sjekk om tilpasningen blir vesentlig (signifikant) bedre når du legger til en blokk.
   2. I hierarkisk regresjon blir grupper av uavhengige variabler trinnvis lagt til.
   3. F-testen kan benyttes til å vurdere om et trinn signifikant forbedrer modellen.
2. Skrittvis (stepwise)
   1. En helt automatisert fremgangsmåte å velge prediktorer.
   2. en algoritme som autmatisk velger ut variabler for en regresjonsmodell.
   3. Forward selection: Start uten uavhsngige variabler i modellen. Test hvor mye hver enkelt nye variabel ville forbedret modellen, og legg til den som forbedrer modellen mest. Gjenta denne prosessen til ingen nye variabler signifikant forbedrer modellen.
3. Informasjonskriterier
   1. Bruk statistikker som finner den beste balansen mellom undertilpasning og overtilpasning.
   2. er 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).

* Begrensninger med automatiske fremgangsmåter
  + Stepwise regresjon er en fulstendig automatisert prosess der dataene får drive modellen. Mange misliker at statistiske kriterier får informere teori uten ukritisk vurdering eller teoretisk forankring (det er ingen garanti for at den endelige modellen "gir mening").
  + Det er ingen garanti for at fremover og bakover-prosedyrene vil resultere i at du velger ut det samme settet med uavhengige variabler.
  + Du bruker det samme datasettet både til å velge prediktorer, og vurdere signifikans

**Overtilpasning vs. undertilpasning**:

* 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 dataset.
* Overtilpasning: Dårlig prediksjon som følge av at du lærer for mye fra data.
* Undertilpasning: Dårlig prediksjon som følge av at du lærer for lite fra data.

**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 en hjørnesten i moderne maskinlæringsmetoder.
* 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 tesdataene å synke.

# Forelesning 6: Kjikvadratanalyser

**Kategorisk data**: Data der verdiene utgjør et bestemt antall distinkte kategorier.

* + Nominelle variabler: Ingen rangering av kategoriene (f.eks. bilmerker).
  + Ordinale variabler: Distinkte kategorier, men de har en rangering (f.eks. vektklasser i boksing).
* Regresjon er egnet når den avhengige variabelen er kvantitativ.
* Hva kan du gjøre dersom du er interessert i å se på sammenhengen mellom to kategoriske variabler?

**Stokastisk**: 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.

**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:

**Additiv regel for sannsynlighet**: Dersom A og B er gjensidig utelukkende er sannsynligheten for enten A eller B:

* Hvis A og B er to hendelser, så er sannsynligheten.
* For gjensidig utelukkende sett forenkles dette til

**Simultan sannsynlighet**: Sannsynligheten for at flere utfall inntreffer.

**Betinget sannsynlighet**: For enhver 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:

**Multiplikasjonsregelen**: Dersom A og B er uavhengige utfall, så er den simultane sannsynligheten (sannsynligheten for begge) gitt ved:

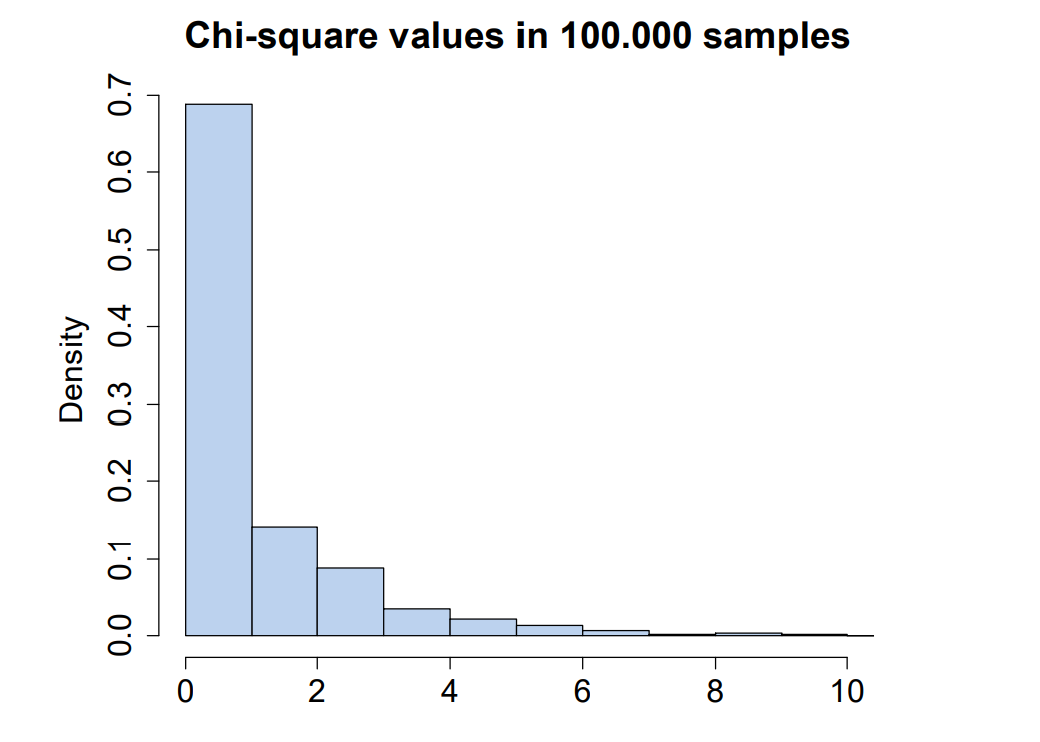
**Kji-kvadrat-tester**:

1. Kji-kvadrat goodness of fit test
   1. Teste i hvilken grad verdiene på en observert variabel er i overenstemmelse med en bestemt fordeling.
2. Kji-kvadrat test av uavhengighet
   1. Test på i hvilken grad to variabler er (statistisk) uavhengige.

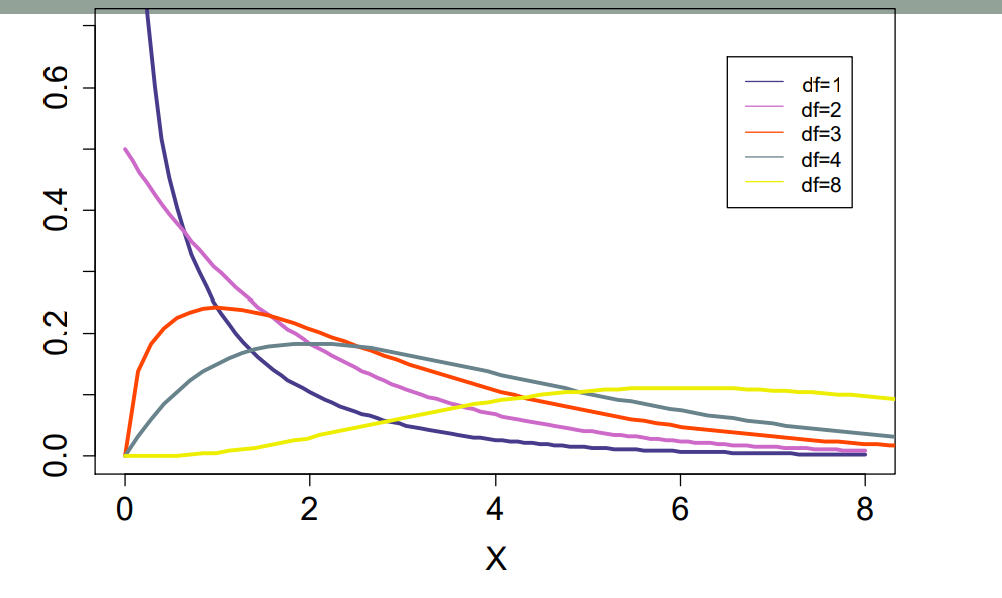
* Skillet viser ulik bruk av kji-kvadrat statistikken. Beregningen er den samme i begge tilfeller.
* Testen på uavhengighet vil være den viktigste for oss.

**Kji-kvadrat-statistikken**:

* Legg merke til:
  + Denne summen blir liten når de observerte frekvensene har verdier som er lik de forventede.
  + Vi deler på forventet antall, fordi en forskjell mellom observert og forventet antall må sees i forhold til hvor mange man forventer.
    - Det er mindre relativ forskjell på 1000og 1001, enn på 10 og 11, selv om forskjellen i begge tilfeller er «1».



* Samplingfordelingen til kji-kvadrat statistikken under H0



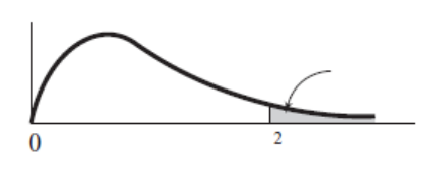
* Under H0 følger den ovserverte verdien for kji-kvadrat statistikken en kji-kvadrat fordeling.
  + Formen på en kji-kvadrat fordeling bestemmes av ett tall (antall frihetsgrader).
  + Vi må altså vite antall frihetsgrader. I vårt eksempel er df=1. Vi kan da finne den kritiske verdien i en kji-kvadrat tabell.

**Kji-kvadrat tabell**:

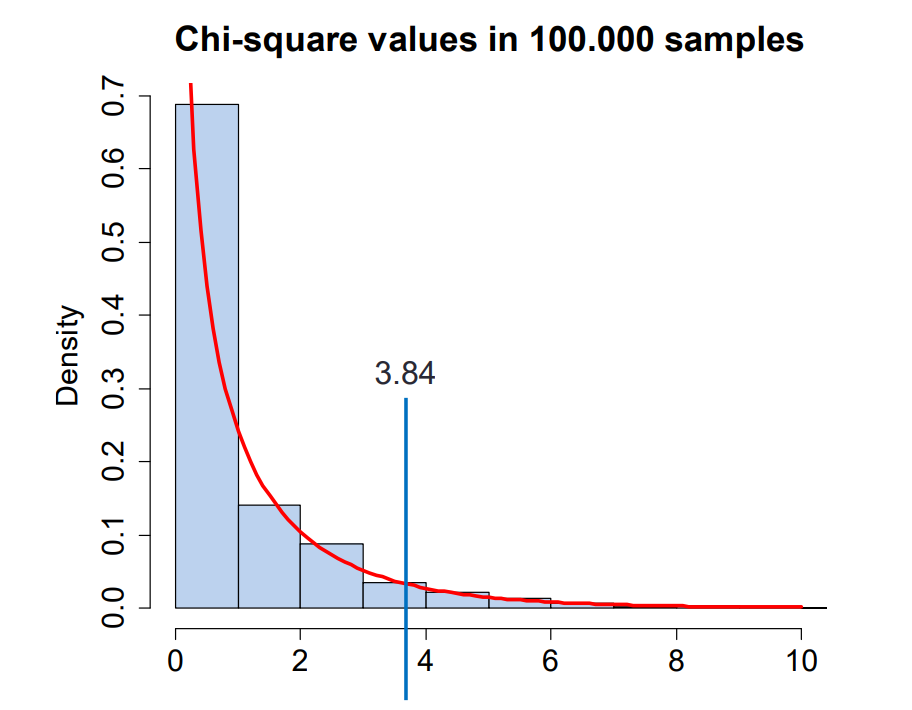
Et bilde som inneholder bord

Automatisk generert beskrivelse

* Den kritiske verdien for en kji-kvadratfordeling med en frihetsgrad er 3.84



**Kji-test distribusjon**:



* The sampling distribution follows a Chi-square distribution, and 95% of the area of this distribution lies to the left of 3.84.

**Analyse av residualene**:

* Sammenlikning av de observerte og forventede frekvensene lar oss teste hypotesen om uavhengighet.
* Forskjellen fo − fe , differansen mellom observer frekvens of forventet frekvens kalles et residual.
* Hvor stort må et residual være før det er av betydning? Også her standardiserer vi residualene.

Et bilde som inneholder tekst, bord

Automatisk generert beskrivelse

**Kji-kvadrat-analyse**:

* Begrenset 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 (typisk kreves det at det forventede antallet er minst 5 i alle cellene).
* For valide resultater kreves det at observasjonene er uavhengige av hverandre.

# Forelesning 7: Variansanalyse

**Dummyvariabler**: 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.

**Testing av parvise gjennomsnitt**: Testen vi har gjort sjekker om det er forskjell mellom noen av gruppene, men ikke mellom hvilke. Vi kan sjekke dette også.

* La være gjennomsnittet i gruppe , være antall observasjoner i gruppe g,
* Og være Residual Mean Square. Da er et konfidensintervall for forskjellen mellom to grupper g1 og g2 gitt ved:

Et bilde som inneholder tekst

Automatisk generert beskrivelse

**Multippel testing**: Når antallet grupper vokser, vokser antallet parvise sammenligninger mye raskere.

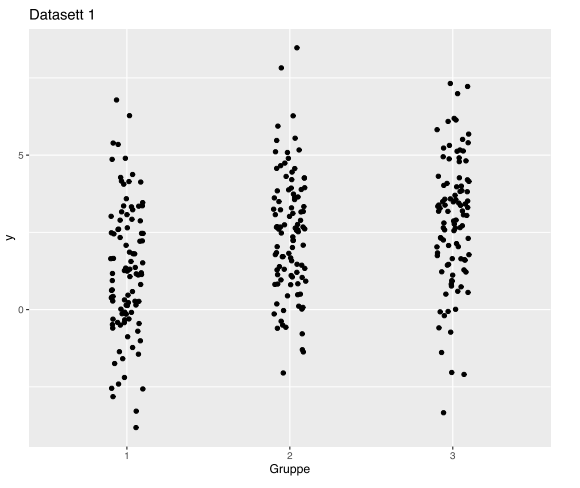
**Bonferroni-korreksjon**: Bonferroni-korreksjon endrer bredden på konfidensintervallene, slik at vi kontrollerer den totale sannsynligheten for å gjøre feil.

**Tukeys metode for multippel korreksjon**: Bonferronis metode er konservativ. Den garanterer 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.

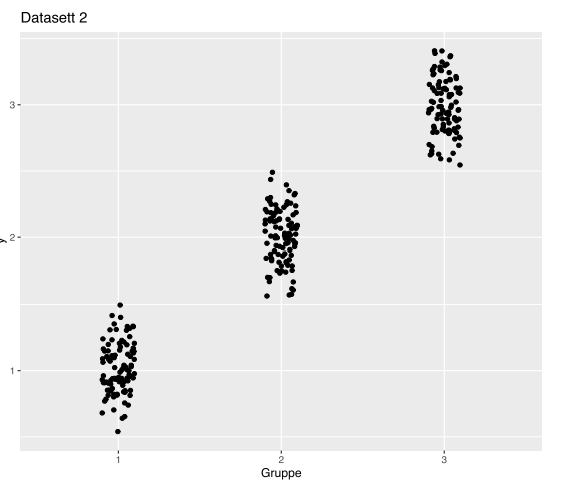


**Within-groups variance**: Variasjon innad i gruppene (within-groups variance) beregnes utifra hvor mye observasjonene i én gruppe varierer fra gjennomsnittet .

* Variasjon innad i gruppene er større enn variasjonen mellom gruppene.



**Between-groups variance**: Variasjonen mellom gruppene (between-groups variance) beregnes utifra hvor mye gjennomsnittene i hver gruppe varierer fra det total gjennomsnittet .

* Variasjonen innad i gruppene er mindre enn variasjonen mellom gruppene. 

**Variansanalyse**: Teststatistikken er:

* 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 i forhold til Varians innad gjør oss sikrere på at det er reell forskjell mellom gruppene.

**Toveis variansanalyse**: I en toveis variansanalyse har vi to uavhengige variabler. Vi kan også utføre treveis, fireveis, osv, men da blir det fort vanskelig å skjønne hva vi holder på med.

* Vi får derfor to teststatistikker, én for hver faktor.
* Telleren er variasjonen mellom gruppene for faktoren.
* Nevneren er variasjonen innenfor grupper.

# Forelesning 8: Mer variansanalyse og bootstrapping

**Huskeregler for interaksjoner**: Dersom interaksjonen er signifikant, men hovedeffektene ikke er det, skal hovedeffektene likevel med i modellen. Interaksjoner gir ikke mening uten at hovedeffektene også er der.

* Interaksjoner betyr at effekten av én variabel varierer basert på verdien av en annen variabel. Begge må derfor tolkes samtidig.

**Kilder til varians**: Variasjon mellom individer og innad i individer.

**Random effects**: Random effects er ikke parametre på samme måte som -ene. De er i stedet ekstra residualledd som lar oss dele inn i variasjon innad og mellom.

**Bootstrapping**: En simuleringsmetode, som fungerer som følger.

1. Bestem et stort tall , f.eks. 10000
2. Trekk rader fra det opprinnelige datasettet, med tilbakelegging, slik at du får nye like store datasett.
3. Tilpass modeller, én til hvert av de nye datasettene, og for hver modell lagre verdien for statistikken av interesse, f.eks, beta4 + beta1.

* Vi har nå en fordeling med verdier av beta4 + beta1. Vi kan bruke prosentilene av denne fordelingen for å finne konfidensintervaller og -verdier.
* Veldig nyttig teknikk når du trenger konfidensintervaller som ikke kommer rett ut av programmet.

**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.