09 - Korrelasjon 2

Forelesningsmål

Lære om regresjon og Spearman rank order korrelasjon.

Observasjon og reliabilitet ble ikke gjennomgått på forelesning, må leses på egenhånd.

Relevante eksamensoppgaver

- *Vår 2022*

  En gruppe forskere finner en interessant sammenheng der 25 % av den norske befolkningen opplever vårslapphet og økt tretthet når det går mot varmere og lysere dager.

  > d. Regresjonsanalyse kan betraktes som en videreføring av korrelasjon. Forklar hvorfor, og bruk de to variablene over som eksempel.

- *Vår 2020*

  Du finner at tid brukt på å spille voldelige dataspill og aggresjonsnivå hos barn ser ut til å ha en sammenheng med hverandre.

  > d) Hva forstår du med delt varians (Coefficient of determination) i denne sammenhengen, og hvor stor er den i dette eksempelet?
  >
  > e) Regresjonsanalyse kan betraktes som en videreføring av korrelasjon. Forklar hvorfor, og bruk gjerne de to variablene over som eksempel.

- *Vår 2019*

  Det ser ut til å være en sammenheng mellom røyking og høyt skolefravær blant ungdom.

  > b) Hva forstår du med delt varians (Coefficient of determination) i denne sammenhengen, og hvor stor er den?
  >
  > d) Regresjonsanalyse kan betraktes som en videreføring av korrelasjon. Forklar hvorfor, og bruk de to variablene over som eksempel.

- *Høst 2017*

  Du finner at tid brukt på voldelige dataspill og aggresjonsnivå hos barn ser ut til å ha en sammenheng med hverandre.

  c) Regresjonsanalyse kan betraktes som en videreføring av korrelasjon. Forklar hvorfor, og bruk gjerne de to variablene over som eksempel.

  d) Hva forstår du med «minste kvadraters metode» (least-squares criterion) i regresjonsanalyse? Eksemplifiser gjerne ved hjelp av en figur.

Begreper o.l

- **korrelasjonskoeffisient(er) -** *et mål for korrelasjon som går fra -1 (100% negativ korrelasjon) til +1 (100% positiv korrelasjon. Det finnes mange forskjellige typer korrelasjonskoeffisienter, feks Pearsons, Spearman, osv.*

  "Styrken av korrelasjonen gis ved korrelasjonskoeffisienten. En korrelasjonskoeffisient er et tall mellom −1 og 1 som oppsummerer graden av samsvar. Et positivt tall betyr at det er positiv korrelasjon, og et negativt tall betyr at det er negativ korrelasjon. For eksempel fant man i en undersøkelse en korrelasjon på 0,37 mellom høyde og vekt for 7000 voksne kvinner.

  Jo nærmere 1 eller −1 tallet er, jo sterkere er sammenhengen mellom de to størrelsene. I en gruppe med 67 voksne kvinner og menn var det for eksempel en korrelasjon på 0,94 mellom den høyden man ble målt til å ha og den høyden man trodde man hadde.

  Korrelasjonskoeffisienter påvirkes ikke av måleenheten. Det betyr at du får samme tall for korrelasjonen hvis du måler høyde og vekt i meter og kilo, som du får hvis du måler det i centimeter og gram."

  (Store Norske Leksikon)

- **dikotome variabler** - *"variabler med bare to verdier (todelte). Eks. kjønn (mann, kvinne)"*

  (UiO)

- **kontinuerlige variabler** - *variabler som kan ha uendelig mange verdier, feks. alder. Man kan si at ens alder er 86 år. Men man kan også si at ens alder er 86 og 6 måneder, eller 86 år, 6 måneder og 6 dager, eller 86 år, 6 måneder, 6 dager, 6 timer osvosvosv. Kan "spisses" i det uendelige.*

- **Point-Biserial-korrelasjon** - *korrelasjonskoeffisient brukt til å måle korrelasjon mellom en dikotom og en kontinuerlig variabel*

  (fra forelesningsslide)

- **Phi- coeffisienten** - *korrelasjonskoeffisient brukt til å måle korrelasjon mellom to dikotome variabler*

  (fra forelesningsslide)

- **målenivå** - *"Målenivå er en egenskap ved variabler i kvantitativ forskning. Målenivået er et uttrykk for hvordan en variabel er inndelt i verdier. Vi skiller mellom fire målenivåer:*

  - *Nominalnivå: På slike variabler har enhetene enten like eller ulike verdier, som for eksempel variabelen kjønn, med verdiene mann og kvinne. (Kvinner er hverken mer eller mindre enn menn, de er bare to forskjellige typer, red.anm.)*

  - *Ordinalnivå: Her er verdiene ordnet i en bestemt rekkefølge (større eller mindre), som for eksempel på variabelen utdanning, med verdiene lav, middels og høy. (Men forskjellen er ikke kvantifisert, forskjellen mellom middels og høy har ingen tallverdi og er ikke det samme som forskjellen mellom middels og lav - red. anm.)*

  - *Intervallnivå: Her er det også faste avstander mellom verdiene, som for eksempel på variabelen temperatur, med grader som verdier, der avstanden mellom 10 og 20 grader er like stor som mellom 20 og 30 grader. Disse verdiene kan adderes og subtraheres: 15 grader er 10 grader mer enn 5 grader og 20 grader mindre enn 35 grader. (Men merk at 20 grader celcius faktisk -ikke- er dobbelt så varmt som 10 grader celsius, det er bare dobbelt så mye på skalaen. Hadde det vært dobbelt så varmt, slikt som 20 grader kelvin er dobbelt så mye som 10 grader kelvin, hadde det vært på forholdstallsnivå - red.anm.)*

  - *Forholdstallsnivå: I tillegg til faste avstander er det her også et bestemt forhold mellom verdiene, som for eksempel på variabelen alder, med antall år som verdier. Her gir det mening å si at en 40-åring er dobbelt så gammel som en 20-åring og halvparten så gammel som en 80-åring.*

  Forskjellen mellom intervallnivå og forholdstallsnivå er at bare variabler på forholdstallsnivå har et fast definert 0-punkt. Mens 0-punktet for alder er fastsatt ved fødselen, kan temperatur-variabler ha ulike definisjoner av 0-punktet, som for eksempel Celsius-skalaen og Fahrenheit-skalaen. I samfunnsvitenskapen er det vanlig at variabler på intervallnivå også oppfyller kravet til forholdstallsnivå.

  Variablenes målenivå er avgjørende for hvilke analysemetoder som kan brukes. For eksempel vil regresjonsanalyse forutsette at variablene er på intervall- eller forholdstallsnivå."

(Store Norske Leksikon)

- **Spearman rank-order-korrelasjon (rho/ρ)** - *korrelasjonskoeffisient brukt til å måle korrelasjon mellom to variabler på ordinalnivå. Også kjent som Spearmans rank order korrelasjon. Ofte brukes symbolet r_s.*

  Fremgangsmåte for å beregne r_s:

  - *Data ordnes fra lavest til høyest på begge variabler (X og Y).*

  - *Datapunktet med laveste verdi på X aksen får indeks 1 (rangering 1), deretter indeks 2 for nest laveste verdi (så 3, 4 osv).*

    Hvis X-verdien feks er utdanning, gis personen med lavest utdanning indeks/rangering 1. Kanskje er dette en person som aldri har gått på skole. Personen med nest lavest utdanning droppet kanskje ut i 5. klasse. Denne får da indeks/rangering 2. Om det er 10 personer i datasettet ditt, får personen med høyest utdanning av de 10 index/rangering 10.

  - *Samme prosedyre for Y aksen.*

  - *Beregn vanlig Pearson korrelasjon mellom de nye X og Y rank order skårene, i stedet for de originale skårene.*

  - *Ved like skårer benyttes gjennomsnittet for plasseringen som skårene ville hatt om de var forskjellige.*

    Hvis vi fortsatt bruker datasettet med 10 datapunkter om utdanning som eksempel, se for deg at det i datasettet er 3 personer som er like høyt utdannet. Disse har alle ph.d. og er de 3 høyest utdannede personene i datasettet. Siden de alle har like høy utdanning, mangler du en måte å tilskrive de en individuell rangering/index.

    Dersom de hadde hatt marginalt forskjellig utdanning, hadde de fortsatt vært de 3 høyest utdannede personene i datasettet, og hadde innehatt rangering/index 8, 9 og 10. Du tar derfor summen av indexene/rangeringen de ville ha fått, deler denne på antallet skårer for å finne gjennomsnitt ((8+9+10)/3 = 27/3 = 9) og gir alle 3 skårene denne indeksen/rangeringen (9).

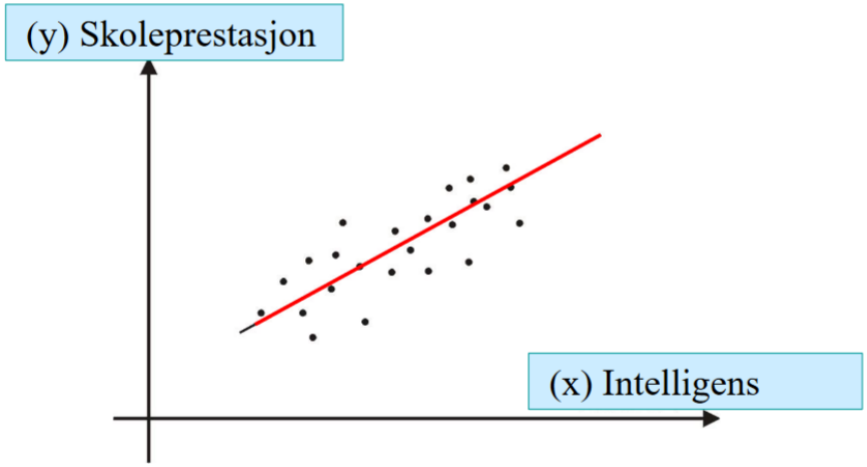
- **regresjon** - *å bruke det du vet om korrelasjonen mellom to eller flere variabler til å predikere verdien på én variabel når den andre variabelen har en viss verdi.*

- **prediktorvariabel (X)** - *variabelen vi predikerer fra*

- **kriterievariabel (Y)** - *variabelen vi predikerer verdien av*

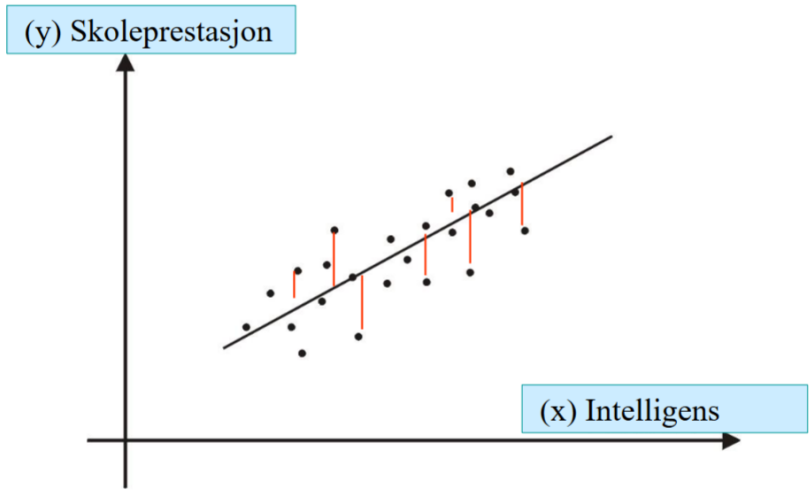
  Ergo: vi predikerer y-verdier fra x-verdier.

- **forklart varians (r^2)** - *andelen av variasjon i y som kan forklares av x. Det finnes to kilder til variasjon, variasjon langs linjen (forklart av x) og variasjon rundt linjen (ikke forklart av x). Når $r^2 = 1$ er all variasjon i y forklart av variasjon i x. I et slikt tilfelle vil samtlige datapunkter ligge på den røde linjen (regresjonslinja).*



- **regresjonslinje** - *den linja som best passer til datapunktene i materialet ditt. Regresjonslinjen gir deg linjen som best minimerer summen av de kvadrerte avstandene mellom hvert punkt og regresjonslinja.*

- **residualer -** *avstand fra datapunkt til regresjonslinjen. Forskjellen mellom predikerte skårer og observerte skårer. «Resten» - det vi ikke har forklart med forklaringsvariabelen gjennom regresjonslinjen. Se røde streker i bildet under.*



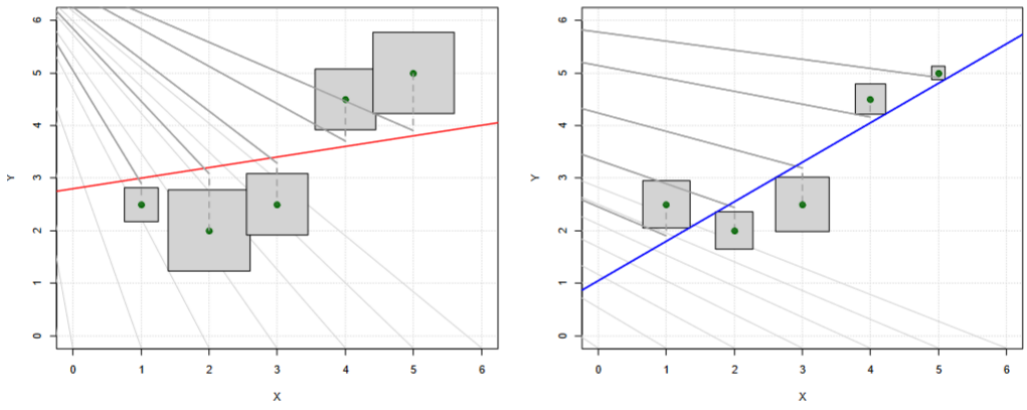
Residual: Forskjellen mellom den observerte og predikerte verdien

$$\hat{e} = Y - \hat{Y}$$

- **standardfeilen av estimatet (SEE - standard error of estimate)** - *et mål på hvor mye den predikerte skåren avviker fra datapunktene.*

- **minste kvadraters metode** - *brukes for å finne linjen som best beskriver sammehengen mellom to variabler. For hver linje kvadreres (ganges med seg selv) residualene før disse (kvadratene av residualene) summeres (legges sammen). Linjen som gir lavest totalsum (sum av kvadratene) er linjen som best passer datapunktene/best beskriver sammenhengen mellom variablene.*

  Fordi residualene kvadreres er metoden sårbar for ekstremer (outliers). Én veldig høy residual vil føre til et enormt kvadrat som vil føre til en veldig høy totalsum. Dermed kan det hende man forkaster en regresjonslinje som overall (for de fleste datapunktene) fungerte veldig godt.



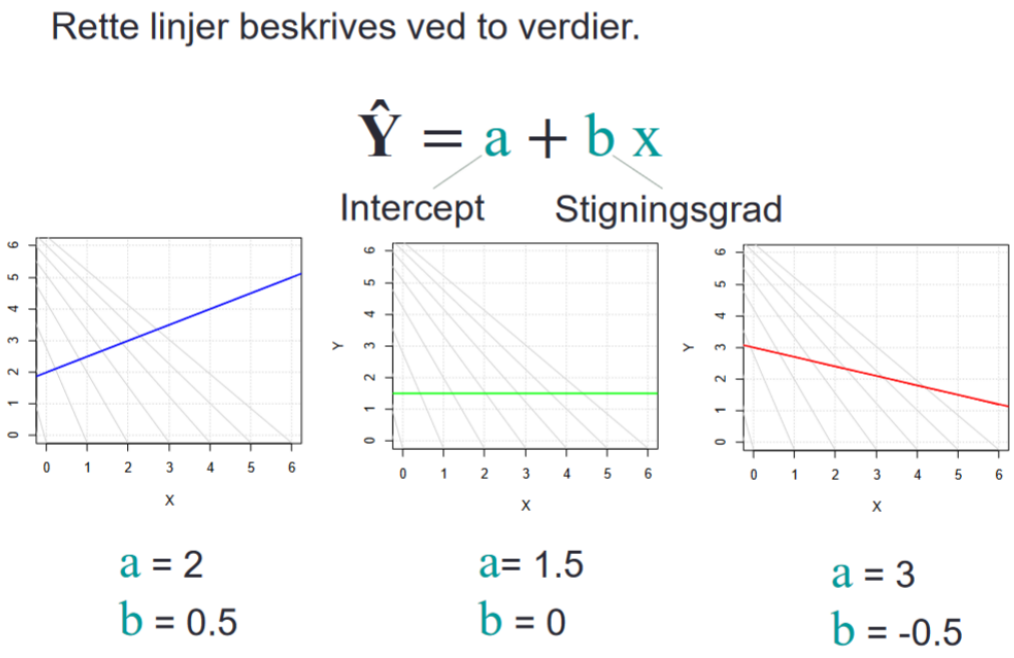
- **lineær regresjon** - *å finne den rette linja som passer best til datapunktene dine.*

  Når vi har funnet denne linjen, og formelen som beskriver den, kan vi anslå en sannsynlig Y-verdi (Ŷ, predikert skåre) for en gitt X-verdi (prediktorvariabel).

  La oss si at du f.eks. har et datasett med høyde og vekt for 10 personer. Alle personene i datasettet er mellom 160cm og 190cm. Ved linær regresjon kan du komme frem til en linje/formel som gir et anslag om sammenhengen mellom høyde og vekt. Du kan så bruke denne formelen for å gjøre et anslag på vekten til en som er 200cm, selv om det ikke var noen som var så høye i datasettet ditt.

  Formel: Ŷ = bX + a

  Ŷ = predikert skåre (kriterievariabel)
  X = variabelen vi predikerer fra (prediktorvariabel)
  b = regresjonskoeffisienten (stigningstallet)
  a = regresjonskonstanten (Y-intercept, punktet der linja krysser Y-aksen)



  For å bestemme Ŷ (predikert skåre) må vi først vite konstantene a og b.

  Dette finner vi ved å først regne ut b.

$$b = \frac{SP = \Sigma(X - \bar{X})(Y - \bar{Y})}{SS_x = \Sigma(X - \bar{X})^2}$$

(Se definisjoner av SP og SS under)

Og så a:

$a = \bar{Y} - b(\bar{X})$

Obs! Merk her at $\bar{Y}$ (gjennomsnittlig Y-verdi, dvs. gjennomsnittet av Y-verdiene vi allerede kjenner) ikke er det samme som $\hat{Y}$ (predikert skåre - Y-verdien vi ønsker å beregne/anslå), og at $\bar{X}$ (gjennomsnittlig X-verdi, dvs. gjennomsnittet av X-verdiene vi allerede kjenner) ikke er det samme som X (en enkelt X-verdi).

Nå kan vi prediker Y-verdien for hvilken som helst skåre på X-variabelen ved å putte tallene inn i regresjonsformelen.

Eksempel: Vi har et datasett bestående av de 8 X- og Y-verdiene i tabellen under (4 X-verdier og 4 Y-verdier). Vi ønsker å predikere Y-verdien når X = 5.

Variabel X: 2, 4, 3, 6 Variabel Y: 2, 5, 1, 4	$\bar{X} = 3,75$ $\bar{Y} = 3,00$
$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
1,75	3,06
0,5	0,063
1,5	0,5
2,25	5
Σ=6 SP	**Σ=8,63 SS_x**

Først finner vi SP og SS_X (se under).

Så finner vi at b = 6 / 8.63 = 0.7 ($b = SP/SS_X$).

Så finner vi at a = 3 - 0.7(3.75) = 0.41 ($a = \bar{Y} - b(\bar{X})$).

Da får vi at $\hat{Y} = 0.7 * 5 + 0.41 = 3.91$ ($\hat{Y} = bX + a$).

- **SP (summen av produkter av avvik/sum of products of deviations)** - *Først tar du den første X-verdien. Så trekker du fra den gjennomsnittlige X-verdien (gjennomsnittet av alle X-verdiene). Så gjør du tilsvarende med den første Y-verdien. Så ganger du disse to resultatene med hverandre. Da har du et produkt (ganget sammen) av to avvik (hvor mye X- og Y-verdiene avviker fra fra hvert sitt gjennomsnitt). Så gjør du dette parvis med alle X- og Y-verdier, før du til slutt summerer alle produktene. Da har du summen av produkter av avvik.*

$$SP_{xy} = \sum(X - \bar{X})(Y - \bar{Y})$$

Eksempel: Den første X-verdien i listen under er 2. Gjennomsnittlig X-verdi ($\bar{X}$) er 3.75. 2 minus 3.75 ($X - \bar{X}$) = -1.75.

Den første Y-verdien er også 2. Gjennomsnittlig Y-verdi ($\bar{Y}$) er 3.
2 minus 3 ($Y - \bar{Y}$) = -1.

*Det første produktet av avvik (produktet av avviket til den første X-verdien og avviket til den første Y-verdien) blir dermed -1.75 * -1 = 1.75.*

Så gjør man det samme for de 3 neste parene med X- og Y-verdier, og summerer alle produktene.

Da finner man at SP = 6.

Variabel X: 2, 4, 3, 6 Variabel Y: 2, 5, 1, 4	$\bar{X} = 3{,}75$ $\bar{Y} = 3{,}00$
$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
1,75 0,5 1,5 2,25	3,06 0,063 0,5 5
Σ=6 SP	**Σ=8,63 SS_X**

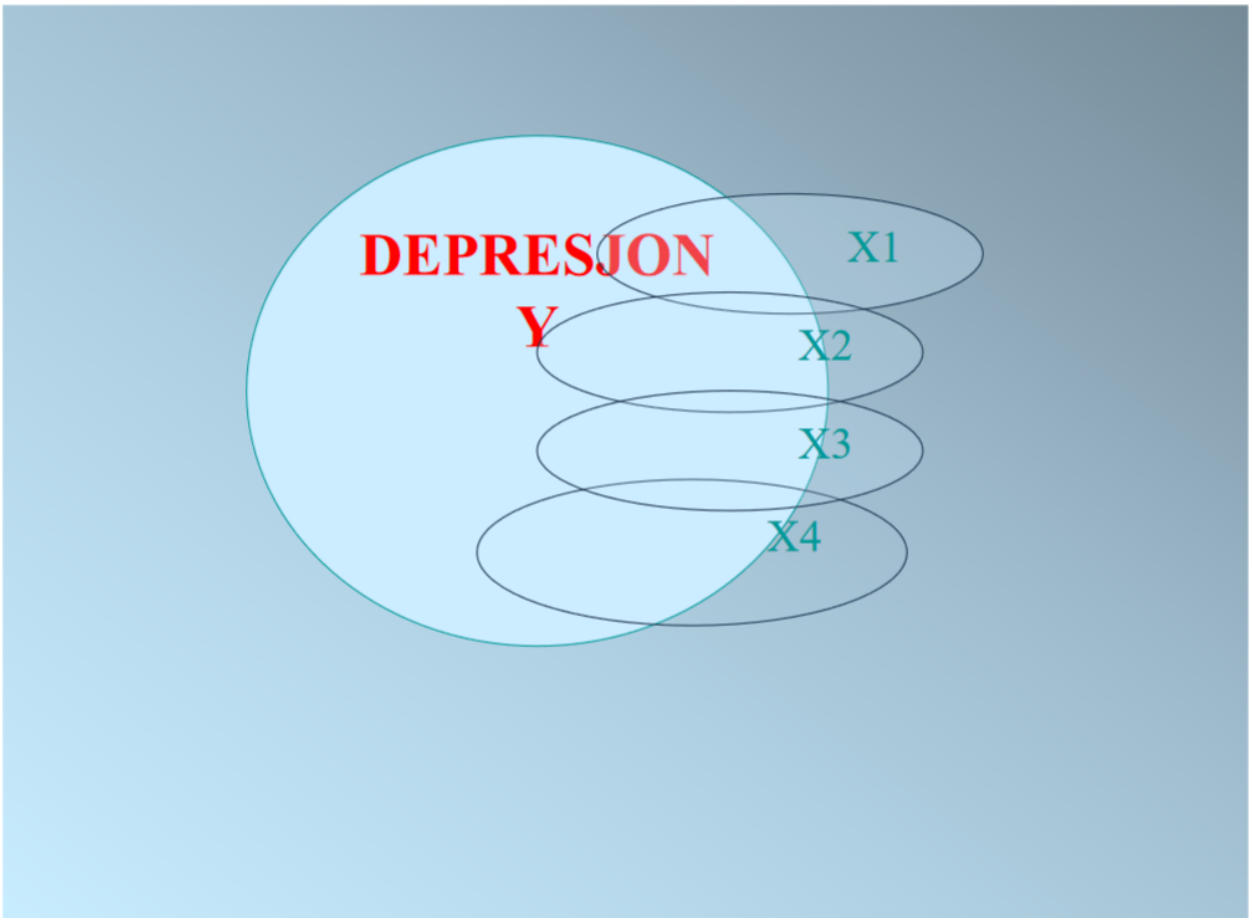
- **SS (sum of squares/summen av kvadrater)** - *Trekk fra gjennomsnittlig verdi fra hver enkelt verdi av den typen og kvadrer (gang med seg selv) resultatene. Summer så alle kvadratene, og du har sum og squares.*

  Kan brukes for X, Y og alle slags andre variabler.

  Eksempel: Den første X-verdien i tabellen over er 2. Gjennomsnittlig X-verdi ($\bar{X}$) er 3.75. $X - \bar{X} = 1.75$. Det første kvadratet ($(X - \bar{X})^2$) blir dermed $1.75 * 1.75 = 3.06$.

  Etter å ha regnet ut de andre kvadratene og summert (Σ) får man at $SS_X = 8.63$.

- **multippel regresjon** - *multippel regresjon er når det er mer enn én prediktorvariabel. For å forklare mest mulig av variansen i kriterievariabelen bør prediktorvariablene være minst mulig korrelerte (think about it!). Ved vanlig multippel regresjon blir alle variablene behandlet likt.*



Et godt egnet scenario for multippel regresjon. Det er lite korrelasjon mellom de ulike X-verdiene (X_1 - X_4) og det vil derfor være mulig å si noe om hvordan de hver for seg korrelerer med/forklarer varians i kriterievariabelen Y (konstruktet vårt - depresjon).

- **hierarkisk regresjon** - *en måte å gjøre regresjon på der man på forhånd antar noe om hvilke variabler som forklarer mest av variansen. Annen måte å regne på.*