

Multippel lineær regresjon

Pål Romundstad



Regresjon

- Regression is the fitting of a function to a set of observations
- Det er minst to variabler involvert som kan deles i to hovedtyper:
 - X-variablene (de uavhengige variablene, kovariater, prediktorer)
 - en respons variabel Y (den avhengige variabelen)
- Y variablene er den observerte responsen til X-variablene slik at de danner parede data punkter

Ett eksempel

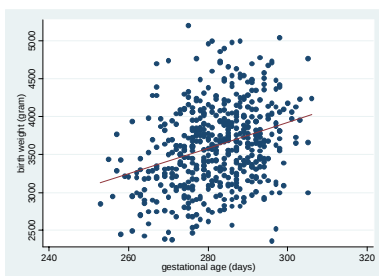
- Gestasjonsalder: Tid mellom starten av svangerskap (eg. "1 dag siste mens") og fødsel
- Normalt=280 dager eller 40 uker
- Hvor mye endrer fødselsvekt seg med økende gestasjonsalder?

Litt data

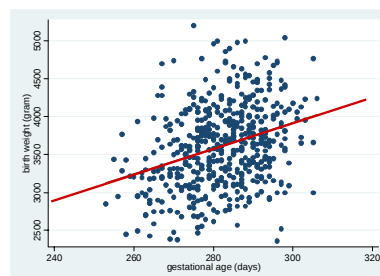
- Målinger av fødselsvekt og gestasjonsalder ved fødsel

Gestasjonsalder	Fødselsvekt
260	3250
272	3505
284	4302
280	3367
276	4200
250	2950
281	3458
265	3260
285	4390
279	3859
293	3670
250	2309
283	3290
280	3987

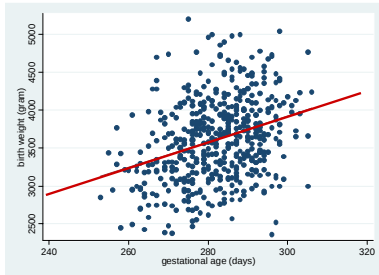
Scatterplot m/ mer data



Ideen: en rett linje



Ideen: en rett linje



modell: $y = a_0 + b_1 x + e$
 y = utkom - avhengig var.
 x = kovariat - uavhengig var.
 e = error, residual

Modell- en rett linje

Kontinuerlig utkom $y = \text{konstant} + \text{regn.koeff} * x + \text{error}$

- Enkel lineær regresjon, en X-variable:

$$y = a_0 + b_1 x + e$$

Multipel lineær regresjon

en teknikk for å estimere den lineære sammenhengen mellom en avhengig variabel og to eller flere uavhengige variabler

Synonym: Multivariabel univariat regresjon

Multivariabel vs multivariat

- **Forklaringsvariabler (X-variabler, uavhengige variabler)**
 - Univariabel- en forklaringsvariabel
 - Multivariabel- flere forklaringsvariabler
- **Utkom-variabler (Y-variabler, avhengige variabler)**
 - Univariat - en utkom variabel
 - Multivariat-flere utkom variabler

Variablene i lineær regresjon

- Y-variabelen: normalfordelt ved gitt X
- X-variablene
 - faste, ikke tilfeldige, gitt på forhånd
 - deres fordeling har ingen betydning for vurdering av usikkerhet
 - (unntak: random-effects models)
- Representasjon av X-er (eks X_1 =alder)
 - kontinuerlig (X_1 = 20 - 80)
 - polynomer: X_1 og X_1^2 etc.
 - dikotom: X_1 = 1 for ung (<40) og X_1 = 0 for gammel (>=50)
 - ordinale kategorier, 10 års-grupper, X_1 = 1,2,3,4,5,6
 - indikatorvariabler (dummy) for hver kategori

Multivariabel lineær regresjon

- ML-regresjon kan ses på som en utvidelse av enkel lineær regresjon til situasjoner der en har to eller flere uavhengige variabler og en avhengig variabel
- To uavhengige variabler: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$
- Generelt :
$$y = \alpha + \sum_{j=1}^k \beta_j x_j + e$$
- Hvor e er normalfordelt med gj.snitt = 0 og varians = σ^2

Multivariabel lineær regresjon

$$y = \alpha + \sum_{j=1}^k \beta_j x_j + e$$

- Parameterene $\alpha, \beta_1, \beta_2, \dots, \beta_k$ estimeres ved minste kvadrats metode
- Residual er avstanden fra en observasjon av Y til regresjonsplanet
- $\beta_j, j = 1, 2, 3, \dots, k$ kalles regresjonskoeffisienter = effektmål
- β_j representerer den predikerte økningen i y per enhets økning i x_j , når alle andre variable holdes konstant (etter justering for alle variablene i modellen) = effektmål
- Hvis det er en sterk sammenheng mellom de uavhengige variablene vil de partielle regresjonskoeffisientene kunne bli svært forskjellig fra koeffisientene en får ved enkel lineær regresjon
- Regresjonskoeffisientene gir et estimat av Y gitt $x_1, x_2, \dots, x_k$

Eksempel-spss

- 2-uavhengige variable-fødselsvekt

Hensikt med multipl lineær regresjon

- **Beskrive sammenhenger - estimering**
 - Estimere sammenheng mellom Y og spesifikk X justert for andre X -variable som mulige konfundere
- **Prediksjon**
 - Prediksjon av Y basert på flere uavhengige variabler (X-er)
 - Innebærer at en generer en modell som bør testes (valideres) i uavhengig datasett

Prediksjon

- Bruke mest mulig bakgrunnsinformasjon
- Bruker x-variabler til å predikere ("gjette") Y-verdier
- Ingen grunn til å spare på antall X-er
- Det bør imidlertid være minst 10 ganger flere observasjoner (målinger av Y-variabelen) enn antall forklaringsvariabler (X-er)
- Modellen evalueres etter prediksjonsstyrke: Hvor god modellen er til å predikere Y, -forklart varians
- La gjerne flere aktuelle modeller konkurrere
- Occam's razor: Av to modeller som passer like godt til dataene er den enkleste den beste

Prediksjon

- Ved prediksjon kan en vurdere å bruke stepwise-prosedyrer:
- Forskjellige variables inkluderes (forward) eller fjernes (backwards) på bakgrunn av hvor betydningsfylle de er i modellen
- I SPSS brukes p-verdi, men bedre å benytte seg av en likelihood ratio test for å vurdere betydning av en variabel i modellen
- Ha kontroll selv
- Prediksjonsmodeller bør evalueres/valideres i uavhengig datasett

Hensikt med multipl lineær regresjon

Beskrive sammenhenger - estimering

- Estimere sammenheng mellom Y og spesifikk X justert for andre X -variable som mulige konfundere
- **Prediksjon**
 - Prediksjon av Y basert på flere uavhengige variabler (X-er)
 - Innebærer at en generer en modell som bør testes (valideres) i uavhengig datasett

Beskrive sammenheng-estimere

- Finne en modell som gir en tolkning av sammenhengen mellom X og Y med justering for andre X-er (kovariater, confoundere)

Justere for konfundere

- Justere for andre uavhengige variable når disse kan betraktes som konfundere
- Effekten av en utenforliggende variabel blander seg med effekten av X-variabelen man ønsker å studere
- En annen uavhengig variabel (en mulig konfunder) er assosiert med både den aktuelle uavhengige variabelen og den avhengige variabelen, men er ikke en del av årsakskjeden mellom aktuelle uavhengige variabelen og den avhengige variabelen
- Er variabelen en del av årsakskjeden (intermediær variabel) kan en risikere å justere bort sammenhengen

Justering-kontrollere for konfundere

- Multivariable modeller
- Mer enn 1 forklaringsvariabel i modellen
- Ved å ha med 2 forklaringsvariable vil effekten av de justeres for hverandre (gjensidig justering av effekt)
- Regresjonen beregner effekten av den ene når den andre holdes konstant og "vice versa"

Beskrive sammenheng

- Bruk Enter-prosedyren (bestem selv hvilke uavhengige variabler som skal inkluderes i modellen)
- Ikke p-verdi som bestemmer hvilke variable det er fornuftig å ha med i modellen, men biologisk kunnskap om sammenhengen som studeres. Modellen bør stemme overens med tidligere erfaring og kunnskap
- Husk!
 - Intermediære variabler i årsaks-kjeden er ikke konfundere
 - Bare ("etablerte") risikofaktorer eller markører for risikofaktorer er reelle confoundere
- X-variable som ikke er signifikante i univariable analyser kan være av betydning i multivariable-analyser

Forutsetninger múltippl lineær regresjon

- Uavhengige residualer (Observasjonene av Y er uavhengige)
- Normalfordelte residualer
- Varians til Y er konstant og lik for alle verdier av X-ene
- Lineær sammenheng mellom X-er og Y

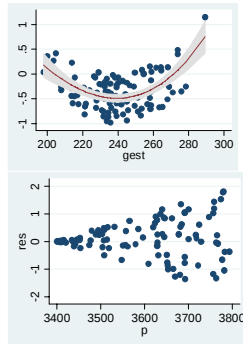
- Avvik fra modellantagelse kan medføre feilaktige konklusjoner

Sjekking av forutsetninger múltippl lineær regresjon

- Scatterplot (X mot Y) for aktuelle variabler (lineær itetsantagelse, konstant varians)
- Residualplot (Y observert minus Y predikert verdi)
Generelt skal residualet være ukorrelet med andre størrelse og plot skal vise tilfeldig spredning rundt antagelser. Spesielle mønstre indikerer avvik fra antagelser
- Normalitetsplott av residual

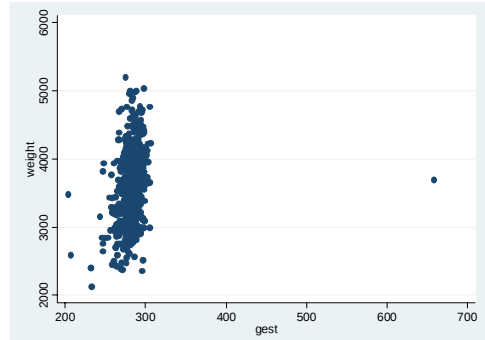
Violations of assumptions

- Non linear effects
Transform the X variable

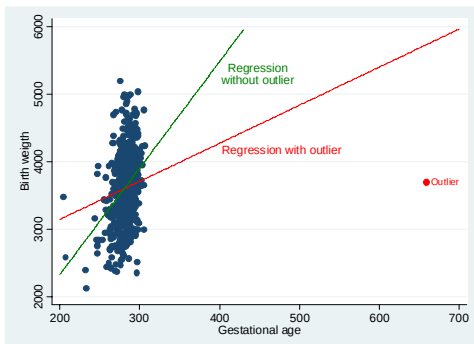


- Non-constant variance
- Use robust SE

Scatterplot



Influence



Prosedyrer ved avvik fra MLR-forutsetninger

- Avvik fra normalfordeling for residualer
 - transformasjon av Y variabel aktuelt (vanligst: $\ln$ transformasjon ved lange haler mot høyre)
 - Andre transformasjoner: $1/x$, $\sqrt{x}$, $1/1-x$, $\cos$
 - Robust SE
- Avvik fra linearitetsantagelse
 - transformasjon av X/Y variabel aktuelt
 - kategorisere X ("dummy" variable som indikatorvariabel)
 - Polynomisk regresjon $y = \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$
- Ved transformasjon endres tolkning av regresjonskoeffisientene da skalaen endres
- Ikke alltid at transformasjon gir bedre tilnærming

Transformasjoner

• $Y = a_0 + a_1 X_1 + a_2 X_2$, Y kan gå fra $-\infty$ til $+\infty$

$\ln(Y) = a_0 + a_1 X + a_2 Z$ som tilsvarer $Y = e^{(a_0 + a_1 X + a_2 Z)}$

Her kan Y bare være et positivt tall (>0)
Modellen går fra additiv til multiplikativ

Additive vs. multiplikative modeller

Additiv modell: $Y = a_0 + a_1 X_1 + a_2 X_2$

• Ved $\ln$ transformasjon av Y får vi:

Multiplikativ modell: $\ln(Y) = a_0 + a_1 X + a_2 X_2$

$Y = e^{(a_0 + a_1 X_1 + a_2 X_2)} = e^{a_0} \cdot e^{a_1 X_1} \cdot e^{a_2 X_2}$

Regresjonskoeffisienten representerer prosentvis økning i Y

Sterkt korrelerte X-variable-kolinearitet

- Ved sterkt korrelerte uavhengige variable er det problematisk å inkludere de i samme modell (regr.koeff. blir dårlig estimert, standardfeilen og p-verdien stiger)
- Før en putter uavhengige variable inn i modellen bør en derfor sjekke korrelasjonen mellom de uavhengige variablene (Korrelasjonsmatrise)
- Ved sterk korrelasjon mellom X-variable, bør en bare inkludere en av disse i modellen (velge ut en)

Forklart varians, R^2

- R^2 indikerer andelen av variansen (variasjonen) i y som kan forklares av regresjonen
- $R^2 \approx 1$ indikerer at all variasjonen forklares av regresjonen
- Hvis en har mange uavhengige variable i modellen og få observasjoner, vil en kunne få en kunstig høy R^2 (såkalt mettet modell)
- Tommelfingerregel: antall observasjoner bør være 10 ganger flere en antall uavhengige variable i modellen
- Justert R^2 tar hensyn til antall uavhengige variable
- Modellen kan være god selv om R^2 er lav

Dummy variable

- Hvis x-variable er diskret (nominal) kan en lage såkalt dummy variable eller kategoriserte indikatorvariable
- f.eks nasjonalitet (norsk, spansk, fransk) kan inkluderes i en multivariat-modell ved å lage to nye binære (0,1) variable (den tredje nasjonaliteten er gitt av de to andre når begge disse er 0)
- $Z_{norsk} = 1$ når nasjonalitet er norsk, ellers = 0
- $Z_{spansk} = 1$ når nasjonalitet er spansk, ellers = 0
- fransk er automatisk gitt ved at både norsk og spansk = 0, og fungerer som en referansekategori

Stratifisering

- Stratifiserte analyser er separate analyser av undergrupper i materialet
- Stratifiserte analyser bør foretas ved mistanke om at sammenhengen som studeres mellom en uavhengig variabel og en avhengig variabel kan være forskjellig for ulike nivå av en tredje variable
- β -koeffisientene er forskjellige for ulike nivå av en tredje variabel
- F.eks for hvert kjønn, i forskjellige aldersgrupper osv...

Interaksjon

- Et alternativ til stratifisering når sammenhengen som studeres kan være forskjellig for ulike nivå av en tredje variable, er å lage såkalte interaksjonsledd (produktet av to uavhengige variable)
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 * x_2 + e$
- Interaksjonsleddet beskriver endring av sammenhengen mellom en uavhengig variabel (Y) og en avhengig variabel (x_1) over en tredje variabel (x_2)

Interaksjon

- Et alternativ til stratifisering når sammenhengen som studeres kan være forskjellig for ulike nivå av en tredje variable, er å lage såkalte interaksjonsledd (produktet av to uavhengige variable)
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 * x_2 + e$
- Interaksjonsleddet beskriver endring av sammenhengen mellom en uavhengig variabel (Y) og en avhengig variabel (x_1) over en tredje variabel (x_2)

Workflow

- Scatterplots
- Bivariate analysis (correlations, simple regression)
- Regression
 - Model fitting
 - Cofactors in/out
 - Interactions
 - Test of assumptions
 - Independent errors
 - Linear effects
 - Constant error variance
 - Influence (robustness)

Advarsel

Multipel regresjon er ingen universell vidundermetode

- Vær på vakt ved
 - tidserieanalyser (avhengige residualer/konfounding)
 - repeterte målinger på samme individ (avhengighet)
 - distinkte subgrupper-uteliggere
- Forutsetter kjennskap til fenomenet som studeres
- Beskriver statistiske sammenhenger, ikke nødvendigvis årsakssammenhenger
- Alle modeller er gale, noen er brukbare