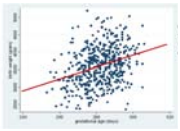
Multippel lineær regresjon

Inger Johanne Bakken
Enhet for anvendt klinisk forskning, NTNU
Og
Avdeling for forebyggende helsearbeid, SINTEF



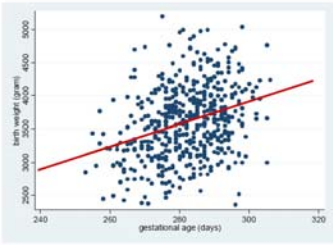
Regresjon

- Tilpasse en funksjon til ett sett observasjoner
- Minst to variable involvert:
  - X (forklaringsvariable, uavhengige variable, kovariater, prediktorer)
  - Y (avhengig variabel, responsvariabel)

Y "hører sammen med" X'ene (parede datapunkter)

Eksempel

- Hvor mye endrer fødselsvekt seg med gestasjonsalder?



Ideen:
Tilpasser en rett linje til datasettet

Modell: En rett linje

- Kontinuerlig utkom y =
  konstant + regresjonskoeffisient * x + støy
- Ved enkel lineær regresjon svarer dette til:

$$y_i = \alpha + \beta^* x_i + e_i$$

Multippel lineær regresjon

- En teknikk for å estimere den lineære sammenhengen mellom én avhengig variabel og to eller flere uavhengige variable.
- Utvidelse av univariabel lineær regresjon
- Synonym: Multivariabel univariat regresjon

Modell:

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

"The reasons for the popularity include the following:

1. A linear model is mathematically simple
2. Its properties are understood with a minimum of statistical theory
3. The components of the model are easy to interpret
4. A linear model appears to be appropriate for a large number of situations
5. A linear model naturally serves as a point of departure for more sophisticated analyses"

Steve Selvin:
Epidemiological analyses
Oxford University Press

OBS!

- Multippel lineær regresjon er ingen vidundermetode
- Alle modeller er gale, noen er brukbare
- Forutsetter kjennskap til fenomenet som studeres
- Beskriver statistiske sammenhenger, ikke nødvendigvis årsakssammenhenger
- Flere observasjoner per person? Avhengige observasjoner og analysemetoden kan ikke benyttes
- Outlayers kan påvirke resultatene mye

Multivariabel eller multivariat? Univariabel eller univariat?

Variabel:	Angår X'ene
Variat:	Angår Y

- Univariabel: én X
- Multivariabel: flere X'er
- Univariat: én Y
- Multivariat: flere Y'er

Multippel lineær regresjon

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

- Hvor e er normalfordelt med forventningsverdi 0 og varians σ^2

Variablene i lineær regresjon

Y: normalfordelte

X: Ingen spesielle krav. Kan ha ulike former

"X-verdiene kan ha ulike former"

Eksempel

Ulike representasjoner av alder:

- kontinuerlig (20, 21, 22 osv)
- dikotom (X=1 for ung, X=0 for gammel)
- ordinale kategorier (1, 2, 3, 4, 5,… angir aldersgruppe, eks. ti-års)
- indikatorverdier (dummy) for hver kategori

Dummyvariable

Eksempel nasjonalitet norsk, fransk eller spansk:

X_{fransk}=1 hvis fransk, 0 ellers
X_{norsk}= 1 hvis norsk, 0 ellers

Trenger ikke egen variabel for "spansk" – "spansk" er gitt hvis både X_{fransk} og X_{norsk} er 0.

Polynom

Kan også være aktuelt representere X_1 som et polynom:

$$\beta_1 * X_1 + \beta_2 * X_1^2 + \beta_3 * X_1^3$$

”Lineær” regresjon?

Modellen er fortsatt gitt av

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

”lineær” i denne sammenhengen henspeiler på at modellen er gitt som en sum (”lineær i β’ene”)

”Lineær” regresjon?

Tabell 11.9 s 511 Rosner

	f_vekt	alder	syst_bp
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88
Total N	16	16	16

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

Hensikt med multippel lineær regresjon

- Beskrive sammenhenger – estimering
  - Estimere sammenhenger mellom Y og spesifikk X justert for andre X-variabler som er mulige konfundere
- Prediksjon
  - Prediksjon av Y basert på flere uavhengige variabler (X-er)
  - Generering av modell som må testes (valideres) i et uavhengig datasett

”Justere”

- Justerer for utenforliggende variabler som blander seg med X-variabelen man ønsker å studere
- Trenger kunnskap om fenomenet for å vurdere hvilke variable som bør inn i modellen. Modellen bør stemme overens med tidligere erfaring og kunnskap
- X-variable som ikke er signifikante i univariable analyser kan være av betydning i multivariable analyser
- Vær oppmerksom på mulig kollinearitet (ved sterk avhengighet mellom to X’er bør modellen ikke inneholde begge)

Sterkt korrelerte X-variable (kollinearitet)

- Ved sterkt korrelerte uavhengige variable er det problematisk å inkludere begge i samme modell (dårlig estimerte regresjonskoeffisienter, høyere standardavvik, høyere p-verdier)
- Sjekk korrelasjon mellom variablene (korrelasjonsmatriser og sunn fornuft)
- Ved sterk korrelasjon: velg ut én av variablene

Stratifisering

- Analyserer ulike strata separat (eks. kjønn)
- Aktuelt ved mistanke om at sammenhengen mellom Y og én X er forskjellig for ulike nivå av en annen X

Interaksjon

- Alternativ til stratifisering
- Inkludere interaksjonsledd i modellen (produktet av to uavhengige variable.
- Med to uavhengige variable blir modellen:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_1 * x_2 + e$$

Forklart varians R^2

- R^2 gir indikasjon for andelen av variansen i den uavhengige variabelen som kan forklares av regresjonen (som i univariabel lineær regresjon)
- R^2=1 indikerer at all variasjonen forklares av regresjonen
- Jo flere forklaringsvariable, jo høyere R^2, må derfor være forsiktig med å tolke R^2. "Mettet modell" – modell med for mange forklaringsvariable i forhold til n
- Justert R^2 tar hensyn til antallet uavhengige variable i modellen

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,939[a]	,881	,863	2,479

a. Predictors: (Constant), alder, f_vekt

Transformasjoner

- Kan være aktuelt ved avvik fra normalfordeling
  - Vanligst: ln transformasjon
- Aktuelt ved avvik fra antagelse om linearitet:
  - Kategorisere X (dummyvariabel)
  - Polynomisk regresjon: $a_1X_1+a_1X_1^2+a_3X_1^3$
- Ved transformasjon endres tolkning av regresjonskoeffisientene (annen skala)
- Ikke alltid at transformasjon gir bedre tilnærming

Om nedre deteksjonsgrense

ln(0) er ikke definert. Hva hvis måleverdien faktisk er 0?

"... There are no simple solutions to this difficulty. Ad hoc solutions, such as replacing zero values by the limit of detection (or perhaps half the limit of detection) if the limit is known, may prove satisfactorily. ..."

Armitage, P., & al: Statistical Methods in Medical Research". 4th Ed, 2002. ISBN 0-632-05257-0 (side 311)

Ln transformasjon av Y: Fra additiv til multiplikativ modell

- Additiv modell med to variable:

$$Y = a_0 + a_1 * X_1 + a_2 * X_2$$

- Multiplikativ modell – ln transformasjon av y

$$\ln(Y) = a_0 + a_1 * X_1 + a_2 * X_2$$

$$Y = \exp(a_0 + a_1 * X_1 + a_2 * X_2)$$

$$= \exp(a) * \exp(a_1 X_1) * \exp(a_2 X_2)$$

Additiv versus multiplikativ modell

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

  a_1=0,007 => Y øker med 0,007 ved én enhets økning i X_1 når X_2 holdes konstant

  a_1=0,800 => Y øker med 0,800 ved én enhets økning i X_1 når X_2 holdes konstant

- **Multiplikativ modell:** $Y = \exp(a)*\exp(a_1X_1)*\exp(a_2X_2)$

a_1 = 0,007 => Y øker med en *faktor* exp(0,007)=1,007 ved én enhets økning i X_1 når X_2 holdes konstant

a_1 = 0,800 => Y øker med en *faktor* exp(0,800)=2,2 ved én enhets økning i X_1 når X_2 holdes konstant

Arbeidsflyt

- Definer hypotesen
- Deskriptiv statistikk!
- Vurder hvilke bakgrunnsvariable som bør inn i modellen ut fra kunnskap om fenomenet (Tommelfingerregel: antall X'er bør ikke være mer enn 10 % av n)
- Sjekk forutsetninger
  - normalfordelt Y med konstant varians for alle X
  - lineære sammenhenger X og Y
- Modellering
- Sjekk modellen(e):
  - Er residualene uavhengige og normalfordelte?
  - Hvilken modell er "best"? (alle modeller er gale, noen er brukbare)
  - Er modellen "robust" (se etter outlayers og sjekk hvordan disse påvirker koeffisientene i modellen)

Eksempel fra Rosner

Tabell 11.9 s 511 Rosner

	f_vekt	alder	syst_bp
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88
Total N	16	16	16

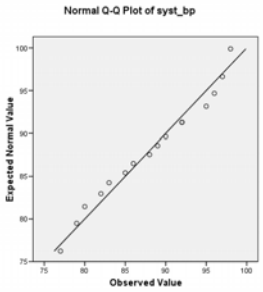
En Y: syst_bp

To X'er: f_vekt (i oz) og alder (i dager)

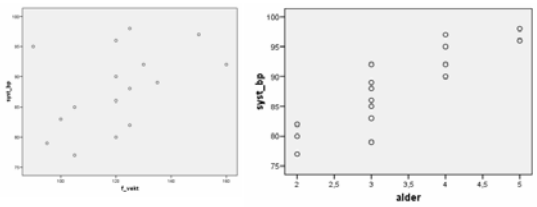
Problemstilling: Hvordan påvirker fødselsvekt og alder systolisk blodtrykk hos nyfødte?

Analysemetode:
Multippel lineær regresjon

Normalfordeling Y?



Lineær sammenheng X'er og Y?



Univariable analyser

Analyze -> regression -> linear regression

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	67,679	3,191		21,212	,000	60,836	74,522
	alder	6,153	,928	,871	6,629	,000	4,162	8,145

a. Dependent Variable: syst_bp

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	69,133	10,410		6,641	,000	46,806	91,460
	f_vekt	,157	,086	,441	1,839	,087	-,026	,341

a. Dependent Variable: syst_bp

Multivariabel analyse

Analyze -> regression -> linear regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	53,450	4,532		11,794	,000	43,660	63,241
	alder	5,888	,680	,833	8,656	,000	4,418	7,357
	f_vekt	,126	,034	,352	3,657	,003	,051	,200

a. Dependent Variable: syst_bp

a. Dependent Variable: syst_bp

$B_{f\_vekt} >> B_{alder}$

Betyr dette at f_vekt har mye større betydning enn alder?

Standardiserte koeffisienter

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	53,450	4,532		11,794	,000	43,660	63,241
	alder	5,888	,680	,833	8,656	,000	4,418	7,357
	f_vekt	,126	,034	,352	3,657	,003	,051	,200

a. Dependent Variable: syst_bp

Standardiserte regresjonskoeffisienter gitt ved

$$b_s = b * (s_x/s_y)$$

Gjennomsnittlig økning i y (uttrykt i standardavvik enheter av y) per standardavvik enhet økning i x, med alle de andre variablene i modellen holdt konstant.

Hypotesetesting

Tester :

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$ mot

H_1: minst én $\beta j \neq 0$

Forkaster H_0 hvis $F > F_{k,n-k-1,1-\alpha}$

hvor a = signifikansnivået
k = antall koeffisienter i modellen
n = antall observasjoner

og F = Reg MS / Res MS

Sum of squares SS

Total SS = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Res SS = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ hvor $\hat{y}_i = a + \sum_{j=1}^{k} b_i x_{ij}$

Reg SS = Total SS - Res SS

Mean squares MS

Reg MS = Reg SS / (n-k-1)

Res MS = Reg SS / k

The test statistic *F*

F = Reg MS / Reg MS

Under H_0 er F F-fordelt med k og (n-k-1) frihetsgrader

Eksempel fra Rosner

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	591,036	2	295,518	48,081	,000[a]
	Residual	79,902	13	6,146		
	Total	670,938	15			

a. Predictors: (Constant), f_vekt, alder

b. Dependent Variable: syst_bp

Tester for bidrag fra hver variabel

To tester (ekvivalente)

1) *t*-tester med $t = b_j/se(b_j)$

Under H_0 (t=0) følger t en t-fordeling med n-k-1 frihetsgrader

2) F-test med

$F =$ $(\text{Reg SS}_{\text{full modell}} - \text{Reg SS}_{\text{alle variablene i modellen unntatt j}}) / \text{Res MS}_{\text{full modell}}$

Under H_0 følger F en F-fordeling med 1 og n-k-1 frihetsgrader

t-test av bidrag fra hver variabel

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	53,450	4,532		11,794	,000	43,660	63,241
	alder	5,888	,680	,833	8,656	,000	4,418	7,357
	f_vekt	,126	,034	,352	3,657	,003	,051	,200

a. Dependent Variable: syst_bp

SPSS gir verdier fra t-testene i tabellen, samt konfidensintervall for parametrene

F-test av variabel f_vekt

$F = (\text{Reg SS}_{\text{full modell}} - \text{Reg SS}_{\text{alle variablene i modellen unntatt f\_vekt}}) / \text{Res MS}_{\text{full modell}}$

Under H_0 følger F en F-fordeling med 1 og n-k-1=13 frihetsgrader

ANOVA[b]

Model		Sum of Squares	df	Mean Square
1	Regression	508,817	1	508,817
	Residual	162,121	14	11,580
	Total	670,938	15	

a. Predictors: (Constant), alder

b. Dependent Variable: syst_bp

ANOVA[b]

Model		Sum of Squares	df	Mean Square
1	Regression	591,036	2	295,518
	Residual	79,902	13	6,146
	Total	670,938	15	

a. Predictors: (Constant), f_vekt, alder

b. Dependent Variable: syst_bp

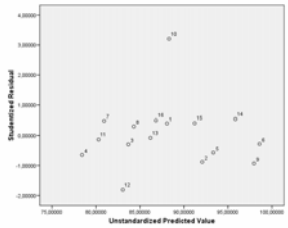
F = (591,0 – 508,8)/6,1= 13,5

Fra Excel: p-verdi = FFdist(13,5;1,13) = 0,003

Sjekk av modellen

Analysis - > regression - > linear regression
Save Predicted values unstandardized (1)
Save Residuals Studentized (2)

Scatterplot av (1) mot (2) (tilsvarende fig 11.23 a)



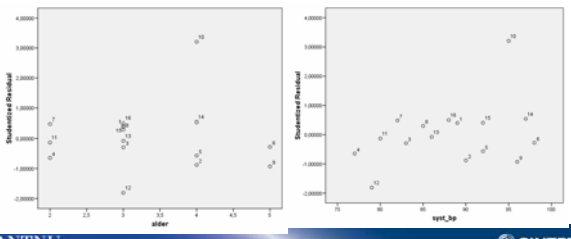
Brukes til å vurdere linearitetsantagelsen og til å se etter outlayers

Ett punkt skiller seg ut

Sjekk av modellen

Analysis - > regression - > linear regression
Save Predicted values unstandardized (1)
Save Residuals Studentized (2)

Scatterplot av (2) mot de to variablene i modellen (tilsvarende fig 11.23 b,c)



Sjekk av modellen

Robust modell?

Hvordan blir modellen uten punkt 10?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	53,450	4,532		11,794	,000
	f_vekt	,126	,034	,352	3,657	,003
	alder	5,888	,680	,833	8,656	,000

a. Dependent Variable: syst_bp

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	47,938	2,302		20,829	,000
	f_vekt	,183	,018	,482	9,955	,000
	alder	5,282	,335	,763	15,759	,000

a. Dependent Variable: syst_bp

Oppsummering

Arbeidsflyt:

- Planlegging, inspeksjon av datasettet
- Normalfordelt Y
  - Transformasjoner?
- Scatterplots (Y mot hver X)
  - Linearitet plausibelt?
- Analyser for hver X (korrelasjoner, univariat regresjon)
- Regresjon
  - Modelltilpasning (forklaringsvariable inn/ut, interaksjon)
- Teste forutsetninger
  - Uavhengige e'er?
  - Lineære effekter?
- Robusthet?

Andre forutsetninger:

- Kunnskap om fenomenet og sunn fornuft

NB! "Alle modeller er gale men noen er brukbare"