

Regression

Morten Frydenberg ©
Institut for Biostatistik

Simpel lineær regression

Data

Model

Antagelser

Parametre

Estimation

Tolkning af output

Modelkontrol

Multipel regression

Korrelation

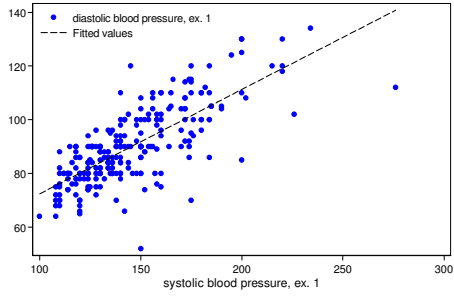
Morten Frydenberg

MPH, Specialmodul - Fælles - Uge 5

1

Lineær regression

Systolisk og diastolisk blodtryk ved først undersøgelse i Frammingham studiet for ikke-rygende mænd:



graph twoway ///
(scatter dbp sbp if (sex==1) & (cig==0), mco(blue) ms(0)) ///
(lfit dbp sbp if (sex==1) & (cig==0)) , legend(ring(0) pos(11))

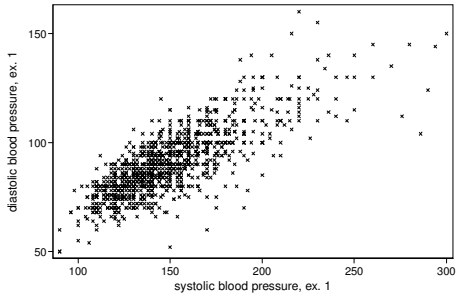
Morten Frydenberg

MPH, Specialmodul - Fælles - Uge 5

4

Lineær regression

Vi ønsker at beskrive sammenhængen mellem **systolisk (SBP)** og **diastolisk (DBP)** blodtryk ved først undersøgelse i Frammingham studiet.



scatter dbp sbp, mco(black) ms(x)

Morten Frydenberg

MPH, Specialmodul - Fælles - Uge 5

2

Den lineære regressions model

Data: n talpar $(x_1, y_1), \dots, (x_n, y_n)$

x Den uafhængige variabel (dosis, eksponering)

y Den afhængige variabel (response)

Modellen: $y = \alpha + \beta \cdot x + \text{error}$

Antagelser:

1. Den forventede værdi af y afhænger lineært af x

2. Afvigelse(rne) (*errors*) er indbyrdes uafhængige

3. Afvigelse(rne) har samme fordeling for alle personer

4. Afvigelse(rne) er normal-fordelte med middelværdi 0 og en ukendt spredning, σ

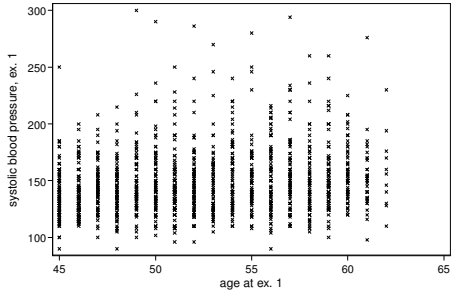
Morten Frydenberg

MPH, Specialmodul - Fælles - Uge 5

5

Lineær regression

Vi ønsker at beskrive sammenhængen mellem **systolisk blodtryk** og **alder** ved først undersøgelse i Frammingham studiet.



scatter sbp age, mco(black) ms(x)

Morten Frydenberg

MPH, Specialmodul - Fælles - Uge 5

3

Den lineære regressions model

$y = \alpha + \beta \cdot x + \text{error}$ $\text{error} \sim N(0, \sigma^2)$

Tre ukendte parametre:

En **hældning** (slope) : β
Tolkning: Forventet forskel i y ved en enhedsforskel i x

En **afskæring** (intercept, constant): α
Tolkning: Forventet værdi af y når x er 0
(Ofte giver dette ingen mening!)

En **spredning omkring linien** : σ
Mål for hvormeget, det ikke kan forklares ved linien.
Et **95% prædiktionsinterval** omkring linien : $\pm 1.96 \cdot \sigma$

Morten Frydenberg

MPH, Specialmodul - Fælles - Uge 5

6

MPH, Fællesmodul - Uge 9

1

Lineær regression: Estimation

$$dbp = \alpha + \beta \cdot sbp + error \quad error \sim N(0, \sigma^2)$$

Estimater findes ved hjælp af en computer - bortset fra mindre forskelle i layout, så vil output blive som dette fra STATA:

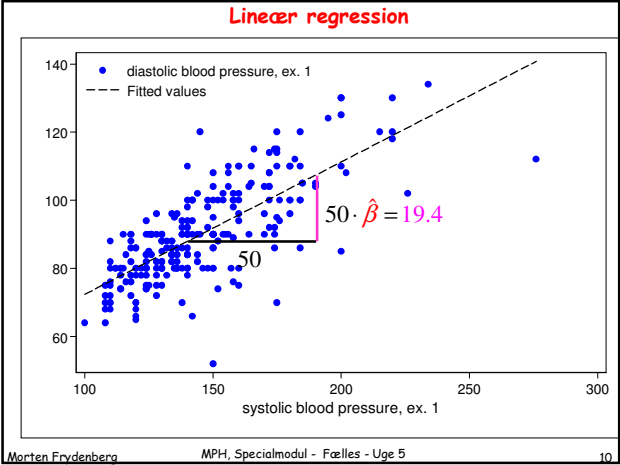
```
regress dbp sbp if (sex==1) & (cig==0)
```

Source	SS	df	MS	Number of obs =
Model	27352.6387	1	27352.6387	256
Residual	21582.9707	254	84.9723254	F(1, 254) = 321.90
Total	48935.6094	255	191.90435	Prob > F = 0.0000

dbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sbp	.3878754	.0216188	17.94	0.000	.3453005 .4304503
_cons	33.64713	3.195451	10.53	0.000	27.35417 39.94008

β α σ^2 σ

Morten Frydenberg MPH, Specialmodul - Fælles - Uge 57



Lineær regression: Estimation

$$dbp = \alpha + \beta \cdot sbp + error \quad error \sim N(0, \sigma^2)$$

I outputtet finder vi også:

Sikkerhedsintervaller for hældning og afskæring.

dbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sbp	.3878754	.0216188	17.94	0.000	.3453005 .4304503
_cons	33.64713	3.195451	10.53	0.000	27.35417 39.94008

Standard error for hældning og afskæring

Test for hypotesen: Hældningen (β) er 0

Morten Frydenberg MPH, Specialmodul - Fælles - Uge 58

Lineær regression: Estimation

$$dbp = \alpha + \beta \cdot sbp + error \quad error \sim N(0, \sigma^2)$$

β : 0.388 (0.345;0.430) Hyp: $\beta=0$ $p<0.001$

α : 33.65 (27.35;39.94) Hyp: $\alpha=0$ $p<0.001$

σ : 9.218

Tolkning af afskæringsestimatet:

Forventet DBP for en person med SBP=0 er 33.6 (27.3;39.9) mmHg.

Dette giver selvfølgelig ingen mening - vi ekstrapolerer!

Tolkning af spredningsestimatet:

95% af observationerne vil ligge mindre end $\pm 1.96 \cdot 9.2 = \pm 18.1$ mmHg fra den estimerede linie.

Morten Frydenberg MPH, Specialmodul - Fælles - Uge 511

Lineær regression: Estimation

$$dbp = \alpha + \beta \cdot sbp + error \quad error \sim N(0, \sigma^2)$$

Estimaterne β : 0.388 (0.345;0.430) Hyp: $\beta=0$ $p<0.001$

α : 33.65 (27.35;39.94) Hyp: $\alpha=0$ $p<0.001$

σ : 9.218

Tolkning af hældningsestimatet:

Hvis to personer afviger med 1 mmHg i SBP, så vil vi forvente at de afviger med 0.388 (0.345;0.430) mmHg i DBP.

Hvis to personer afviger med 5 mmHg i SBP, så vil vi forvente at de afviger med $5 \cdot 0.388$ ($5 \cdot 0.345$; $5 \cdot 0.430$) = 1.94 (1.73;2.15) mmHg i DBP

Data strider imod hypotesen om hældning nul, dvs. hypotesen om ingen sammenhæng mellem SBP og DBP

Morten Frydenberg MPH, Specialmodul - Fælles - Uge 59

Lineær regression: Modelkontrol

$$dbp = \alpha + \beta \cdot sbp + error \quad error \sim N(0, \sigma^2)$$

Modelkontrol indenfor regression består for det meste af at se på plot af residualer:

Residual = Observeret-forventet

$$= y - (\hat{\alpha} + \hat{\beta} \cdot x)$$

I STATA kan man lave en ny variabel, residuals, indeholdende residualerne ved kommandoen:

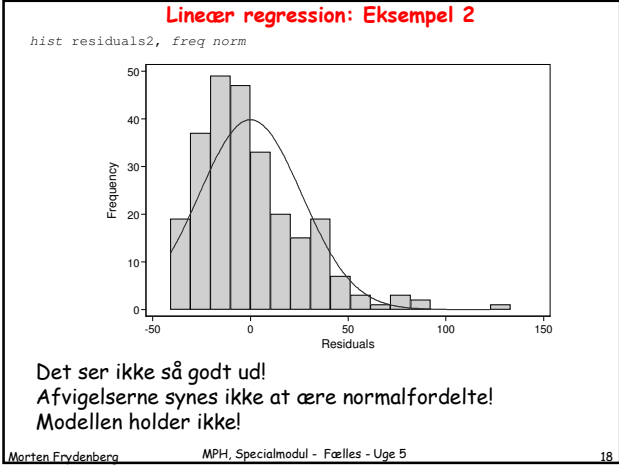
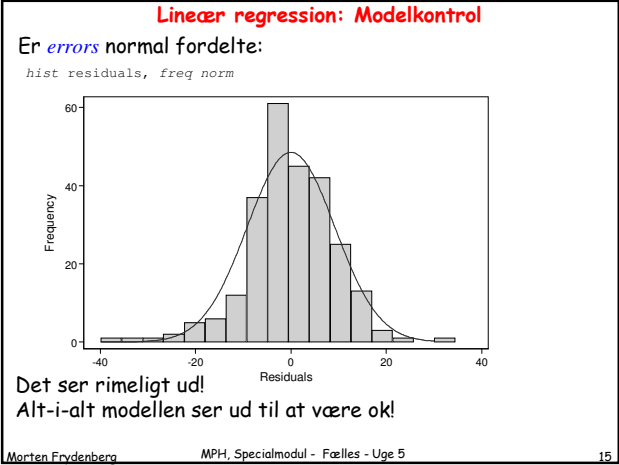
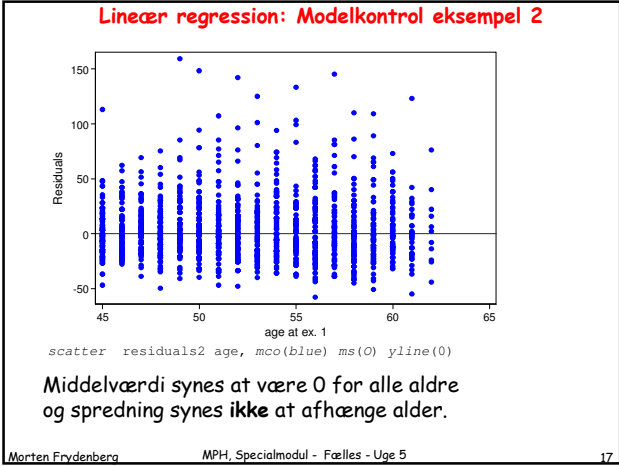
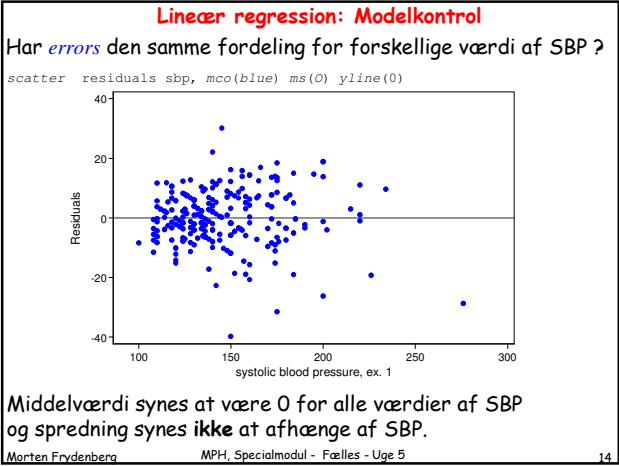
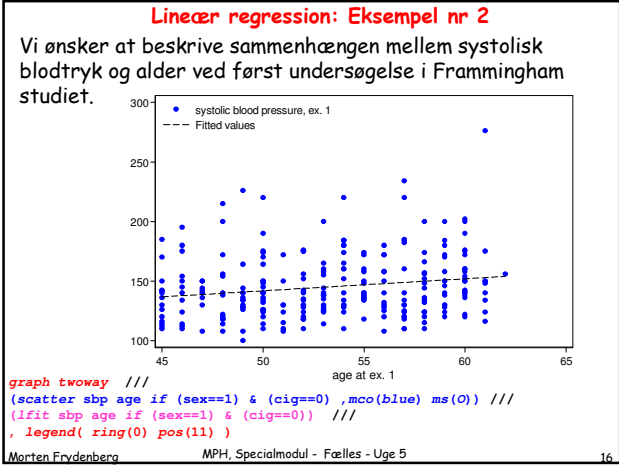
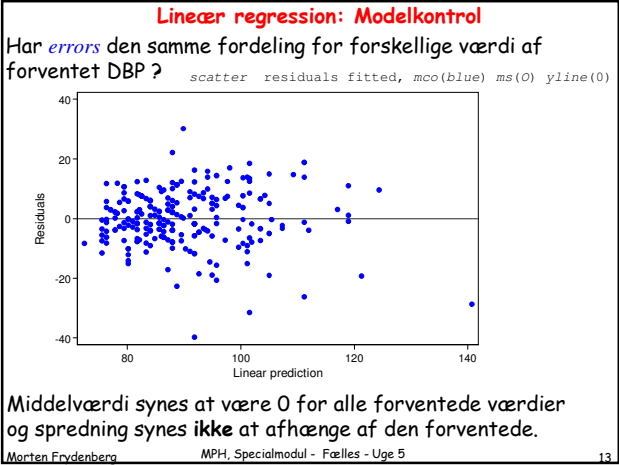
```
predict residuals if e(sample), residual
```

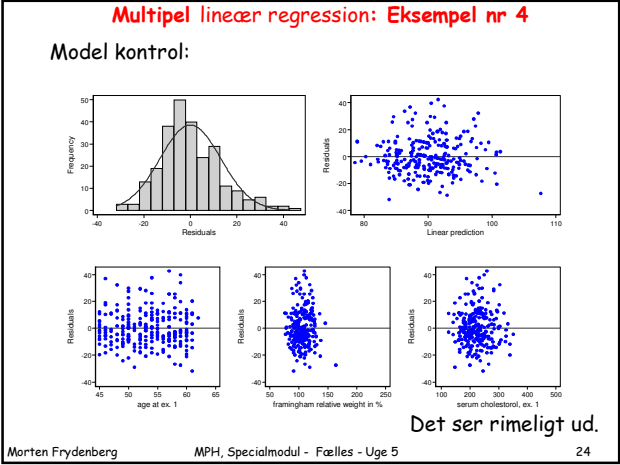
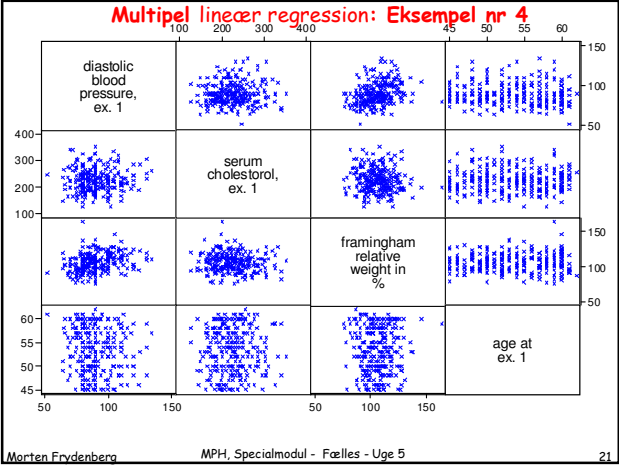
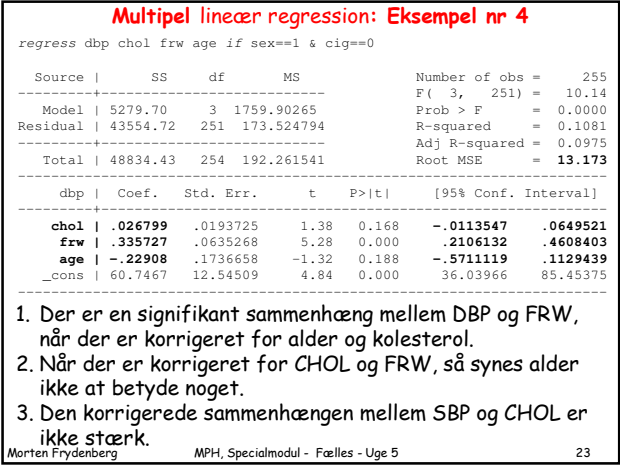
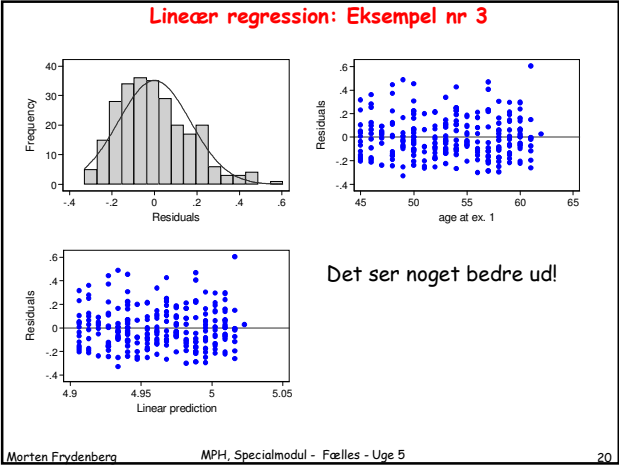
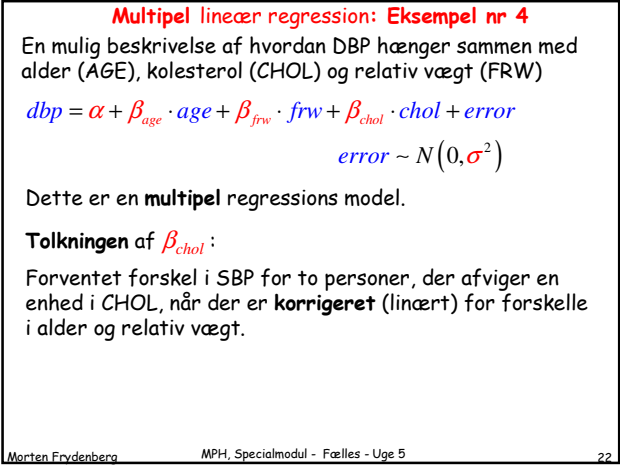
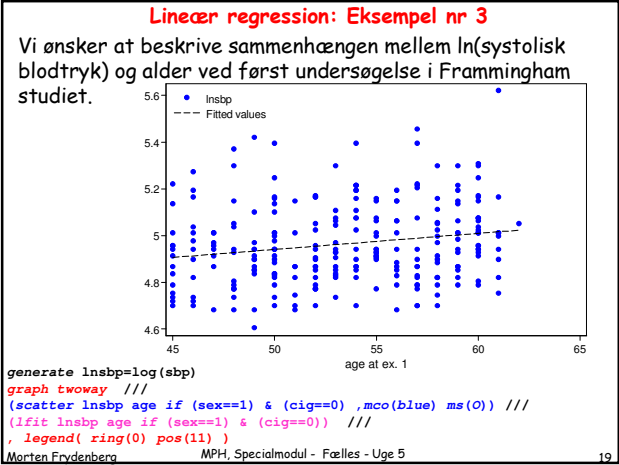
Residualerne er estimater for errors og skal næsten opfylde betingelserne på side.

I STATA kan man lave en ny variabel, fitted, indeholdende residualerne ved kommandoen:

```
predict fitted if e(sample), xb
```

Morten Frydenberg MPH, Specialmodul - Fælles - Uge 512





Lidt om korrelation

Hvis man har **to variable** og ønsker at beskrive sammenhængen mellem disse uden den ene skal antages at være forklaret med den anden, så kan en **Pearson, r , korrelationskoefficient** være svaret.

Vi vil ikke gå i detaljer her, ----

En korrelationskoefficient er et tal mellem -1 og 1:

$$-1 \leq r \leq 1$$

$r=0$, hvis x og y er uafhængige.

$r=1$, hvis og kun hvis præcis lineær sammenhæng med **positiv** hældning

$r=-1$, hvis og kun hvis præcis lineær sammenhæng med **negativ** hældning

Morten FrydenbergMPH, Specialmodul - Fælles - Uge 525

Andre regressionsmodeller

Vi har her set på den situation, hvor den afhængige variabel er **kontinuert** og **forskel middelværdi** det naturlige **associationsmål**.

Hvis man er interesseret i **proportioner** og **odds ratio** er det naturlige associationsmål, så er den mest anvendte modeltype **logistisk regression**.

Hvis man er interesseret i **rater** og **rate ratio** er det naturlige associationsmål, så er de mest **anvendte** modeltyper **poisson regression** eller **cox proportional regression**.

Morten FrydenbergMPH, Specialmodul - Fælles - Uge 528

Lidt om korrelation

Pas på korrelation kan være misvisende.

Her er de ok!

Her ikke er de ok!!!

Lav altid en tegning!!

Morten FrydenbergMPH, Specialmodul - Fælles - Uge 526

Lidt om korrelation

Beregning af korrelation hænger meget sammen med regressionsanalyse.

I det meste output fra regressions analyser finder man r^2 .

Source	SS	df	MS	
Model	27352.6387	1	27352.6387	
Residual	21582.9707	254	84.9723254	
Total	48935.6094	255	191.90435	

Number of obs = 256
F(1, 254) = 321.90
Prob > F = 0.0000
R-squared = 0.5590
Adj R-squared = 0.5572
Root MSE = 9.218

Tolkning af r^2 :

Andel af variation i den ene variabel, der kan forklares ved variation i den anden.

Her: 56% af variation af DBP kan 'forklares' ved variation i SBP og omvendt.

Test for ingen korrelation et ækvivalent med test for $\beta=0$.

Morten FrydenbergMPH, Specialmodul - Fælles - Uge 527