Medisinsk statistikk, del II, vår 2008
KLMED 8005

Eirik Skogvoll
Førsteamanuensis dr. med.
Enhet for Anvendt klinisk forskning
Det medisinske fakultet

1

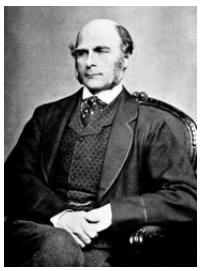
Lineær regresjon, Rosner 11.1 – 11.6

- Bakgrunn (11.1)
- Modell (11.2)
- Estimering av parametre i modellen (11.3)
- Statistisk inferens, testing (11.4)
- Intervallestimering (11.5)
- Modellens godhet

2

”Regresjon…?”

- Betyr ”tilbakegang”
- Sir Francis Galton (1822-1911) observerte at:
  - lange fedre fikk oftest lange barn, men de lengste fedrene fikk i gjennomsnitt kortere barn enn seg selv
  - kortvokste fedre fikk oftest korte barn, men de korteste fedrene fikk i gjennomsnitt lengre barn enn seg selv.



→ ”Regression toward the mean.” A principle stating that of related measurements, and selecting those where the first measurement is either higher or lower than the average, the expected value of the second is closer to the mean than the observed value of the first.

3

11.1 Innledning: østriol vs. fødselsvekt

Observerer (x_i, y_i) for $i = 1, 2, ..., n$

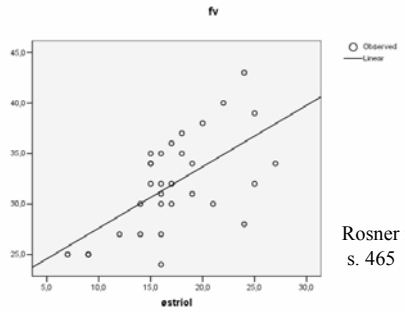
x: prediktor, Y: respons

Rosner tabell 11.1, s. 466

i	østriol	fv
1	7	25
2	9	25
3	9	25
4	12	27
5	14	27
6	16	27
7	16	24
8	14	30
9	16	30
10	16	31
11	17	30
12	19	31
13	21	30
14	24	28
15	15	32
16	16	32
17	17	32
18	25	32
19	27	34
20	15	34
21	15	34
22	15	35
23	16	35
24	19	34
25	18	35
26	17	36
27	18	37
28	20	38
29	22	40
30	25	39
31	24	43

4

"én figur sier mer enn tusen t-tester"



Rosner fig. 11.1, s. 465

5

11.2 Modell

$$Y_i = \alpha + \beta x_i + e_i \quad e_i \sim N(0, \sigma^2)$$

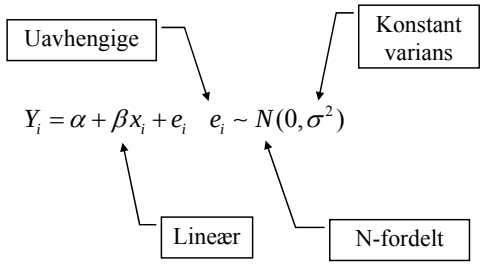
$$E(Y_i \mid x_i) = \alpha + \beta x_i$$

$$Var(Y_i \mid x_i) = \sigma^2$$

$$Y_i \sim N(\alpha + \beta x_i,\ \sigma^2)$$

6

Modellens forutsetninger



7

11.2 Modell (2)

observert: y_i

predikert: $\hat{y}_i = a + bx_i$

gjennomsnitt: $\bar{y}$

8

11.3 Estimering av parametere

- Minste kvadraters metode
- Maximum likelihood

$\hat{\alpha} = a, \quad \hat{\beta} = b$ Rosner fig. 11.4, s. 469

$d_i = y_i - \hat{y}_i = y_i - (a + bx_i)$

$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

9

11.3 Estimering av parametere (2)

observert: y_i

predikert: $\hat{y}_i = a + bx_i$

gjennomsnitt: $\bar{y}$

residual komponent: $y_i - \hat{y}_i$

regresjons-komponent: $\hat{y}_i - \bar{y}$ Rosner fig. 11.5, s. 473

Estimater

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	250,574	1	250,574	17,162	,000[a]
	Residual	423,426	29	14,601		
	Total	674,000	30			

a. Predictors: (Constant), astriol

b. Dependent Variable: fv

$s^2 = \hat{\sigma}^2$

$a = \hat{\alpha}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	21,523	2,620		8,214	,000	16,164	26,883
	astriol	,608	,147	,610	4,143	,000	,308	,908

a. Dependent Variable: fv

$b = \hat{\beta}$

11.4 Inferens: hvor god er modellen?

Rosner fig. 11.6, s. 474:

a) stor regresjonskomponent, liten residualkomponent

b) stor regresjonskomponent, stor residualkomponent

c) liten regresjonskomponent, liten residualkomponent

d) liten regresjonskomponent, stor residualkomponent

best

verst

NTNU

11.4: Hvor god er modellen? (2)

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

Regresjon SS: $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

Residual SS: $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

eller... Total SS = Regresjon SS + Residual SS

13

NTNU

11.4 Hvor god er modellen? (3)
- bedre enn ingen modell?

Regresjon Mean Square: $\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{k}$ k : antall prediktorvariable

(enkel lineær regresjon: $k=1$)

Residual Mean Square: $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k-1}$

(enkel lineær regresjon: $k=1$)

$F=\frac{\text{Reg MS}}{\text{Res MS}} \sim F_{k,n-k-1}$ (enkel lineær regresjon: $F \sim F_{1,n-2}$)

14

NTNU

ANOVA (ANalysis Of VAriance)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	250,574	1	250,574	17,162	,000[a]
	Residual	423,426	29	14,601		
	Total	674,000	30			

a. Predictors: (Constant), østriol

b. Dependent Variable: fv

15

11.4 Hvor god er modellen? (4)

"Coefficient of determination": andel forklart varians:

$$R^2 = \frac{\text{Regresjon SS}}{\text{Total SS}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,610[a]	,372	,350	3,8211	,715

a. Predictors: (Constant), #striol

b. Dependent Variable: fv

- Problem i multippel lineær regresjon: R^2 øker alltid med økende antall prediktorvariable
- R^2 justert: korrigerer for antall prediktorvariable
- Andre, liknende kriterier
  - AIC
  - BIC

16

Hvor god er modellen? (5)

"Compare the models by goodness for purpose".

Professor Brian Ripley, Oxford
(Help-R-forum 17.04.2005)

17

11.4 Tester og konfidensintervall for modellens parametere (koeffisienter)

$\hat{\alpha} = a, \quad \hat{\beta} = b$

$$T = \frac{b}{se(b)} \sim T_{n-2}$$

$$T = \frac{a}{se(a)} \sim T_{n-2}$$

Tosidig $100\% \cdot (1-\alpha)$ CI for β: $b \pm t_{n-2,1-\alpha/2} \cdot se(b)$

Tosidig $100\% \cdot (1-\alpha)$ CI for α: $a \pm t_{n-2,1-\alpha/2} \cdot se(a)$

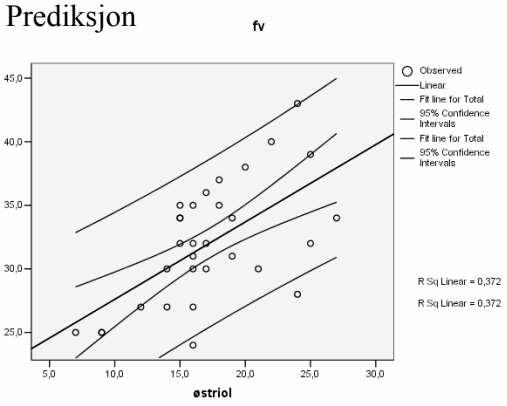
18

Noen sammenhenger…

- I enkel lineær regresjon er hypotesen
  $\beta = 0$ ekvivalent med $\rho = 0$
  som evalueres vhja. Pearsons *r i* korrelasjonsanalysen
- R^2 (”coefficient of determination”) tolkes best i regresjonsanalysen som andelen ”forklart varians” via kvadratsummer.
  Ekvivalent med (Pearsons) r^2

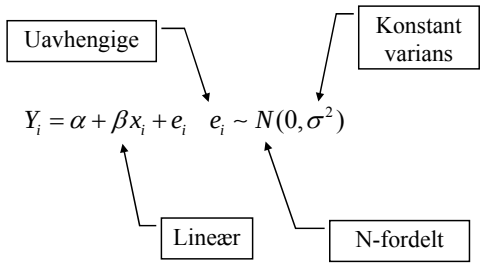
19

Prediksjon



20

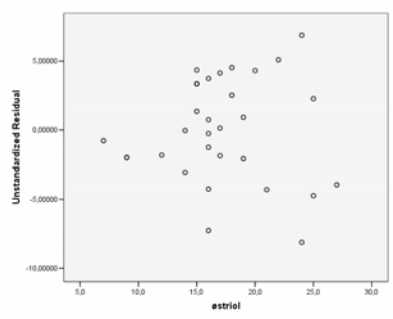
Modellens forutsetninger



21

Sjekk av modellens forutsetninger

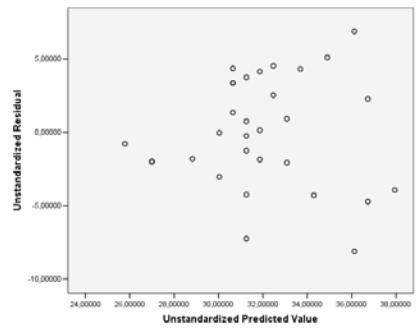
- Linearitet? residual vs. prediktor



22

Sjekk av modellens forutsetninger

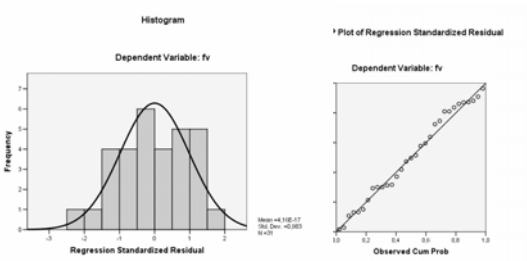
- Konstant varians? residual vs. predikert verdi



23

Sjekk av modellens forutsetninger

-Normalfordelte residualer? histogram og p-p plott



24

Sjekk av modellens forutsetninger
- Uavhengighet mellom feil-leddene

- Relevant for tidsserier (en prosess observert på regelmessige tidspunkter)
- Residual-korrelasjon? Durbin-Watson test