Lineær regression

Læsning:

Jens Ledet Jensen kap. 9

Sammenligning af to middelværdier

Population 1:
Alle studerende
på AU

Population 2:
Alle studerende
på KU

$$\mu_{AU}^{IQ} = \mu_{KU}^{IQ} \, ?$$

Person	IQ
1	110
2	120
3	105
.	.
.	.
.	.
40	140

Person	IQ
1	108
2	118
3	112
.	.
.	.
.	.
55	138

Testkatalog for sammenligning af to middelværdier (<u>u</u>kendt varians)

- Statistisk model
  - $X_{1,i} \sim N(\mu_1, \sigma^2),\ i = 1,2,\dots,n_1$ og $X_{2,i} \sim N(\mu_2, \sigma^2),\ i = 1,2,\dots,n_2$
  - Parameterskøn: $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2)$

    $$s^2 = \frac{1}{n_1+n_2-2}\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)$$
- Hypotesetest
  - $H: \mu_1 = \mu_2$
  - Teststørrelse: $t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$
  - P-værdi: $pval = 2 \cdot \left|1 - t_{cdf}(|t|, n_1 + n_2 - 2)\right|$
- 95% konfidensinterval
  - $[\delta_-; \delta_+] = \left[\bar{x}_1 - \bar{x}_2 - t_0 \cdot s\sqrt{1/n_1 + 1/n_2}\,; \bar{x}_1 - \bar{x}_2 + t_0 \cdot s\sqrt{1/n_1 + 1/n_2}\right]$
  - Hvor $t_0 = t_{inv}(0{,}975, n_1 + n_2 - 2)$

Parrede data

Par nummer	Nyudviklet maskine	Gængs maskine	Forskel
1	8.0	5.6	2.4
2	8.4	7.4	1.0
3	8.0	7.3	0.7
4	6.4	6.4	0.0
5	8.6	7.5	1.1
6	7.7	6.1	1.6
7	7.7	6.6	1.1
8	5.6	6.0	−0.4
9	5.6	5.5	0.1
10	6.2	5.5	0.7

Testkatalog for parrede data

- Statistisk model
  - $d_i = X_{1i} - X_{2i} \sim N(\delta, \sigma^2),\ i = 1,2,\dots,n$
  - Parameterskøn: $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} - x_{2i} \sim N(\delta, \sigma^2/n)$

    $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{1i} - x_{2i})^2$

- Hypotesetest
  - $H: \delta = 0$
  - Teststørrelse: $t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t(n-1)$
  - P-værdi: $pval = 2 \cdot \left|1 - t_{cdf}(|t|, n-1)\right|$

- 95% konfidensinterval
  - $[\delta_-; \delta_+] = \left[\bar{d} - t_0 \cdot s_d\sqrt{1/n}\,; \bar{d} + t_0 \cdot s_d\sqrt{1/n}\right]$
  - Hvor $t_0 = t_{inv}(0{,}975, n-1)$

Parret vs. uparret test

Parret sammenligning: $H: \delta = 0$

Høstudbyttet på 10 marker med nyudviklet maskine.

Høstudbyttet på 10 marker med gængs maskine.

x1(1)	x1(2)	…	x1(10)
x2(1)	x2(2)	…	x2(10)

Differens:

x1(1)-x2(1)	x1(2)-x2(2)	…	x1(10)-x2(10)

→ $\bar{d}$

Her er en eventuel "områdeeffekt" fjernet.

Uparret sammenligning: $H: \mu_1 = \mu_2$

Høstudbyttet på 10 marker med nyudviklet maskine.

x1(1)	x1(2)	…	x1(10)

→ $\bar{x}_1$

Høstudbyttet på 10 marker med gængs maskine.

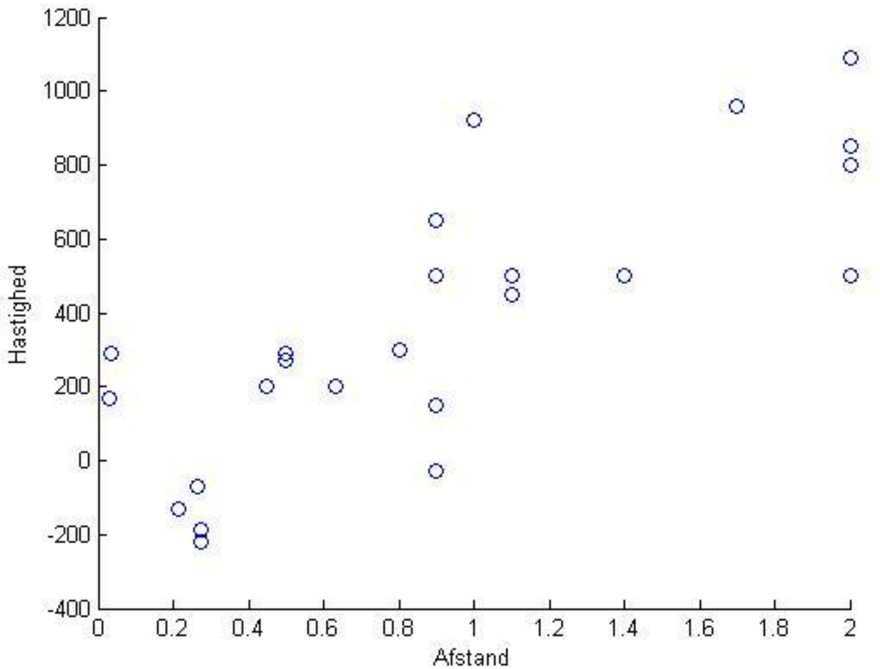
x2(1)	x2(2)	…	x2(10)

→ $\bar{x}_2$

Her er en eventuel "områdeeffekt" ikke fjernet.

Hubbles lov

```
%% Eksempel 1 - Lineær regression - Hubble's målinger
Afstand = [ 0.032 0.034 0.214 0.263 0.275 0.275 0.450 0.500 ...
            0.500 0.630 0.800 0.900 0.900 0.900 0.900 1.000 ...
            1.100 1.100 1.400 1.700 2.000 2.000 2.000 2.000 ];

Hastighed = [ 170 290 -130 -70 -185 -220 200 290 ...
              270 200 300 -30 650 150 500 920 ...
              450 500 500 960 500 850 800 1090 ];
```



Lineær regression

- Hubbles lov og data er et eksempel på en *lineær sammenhæng* mellem
  - en ***responsvariabel*** x og
  - en ***forklarende variabel*** t.
- I Hubbles tilfælde er x hastigheden, hvormed galakserne bevæger sig væk fra hinanden, og t er afstanden mellem galakserne.

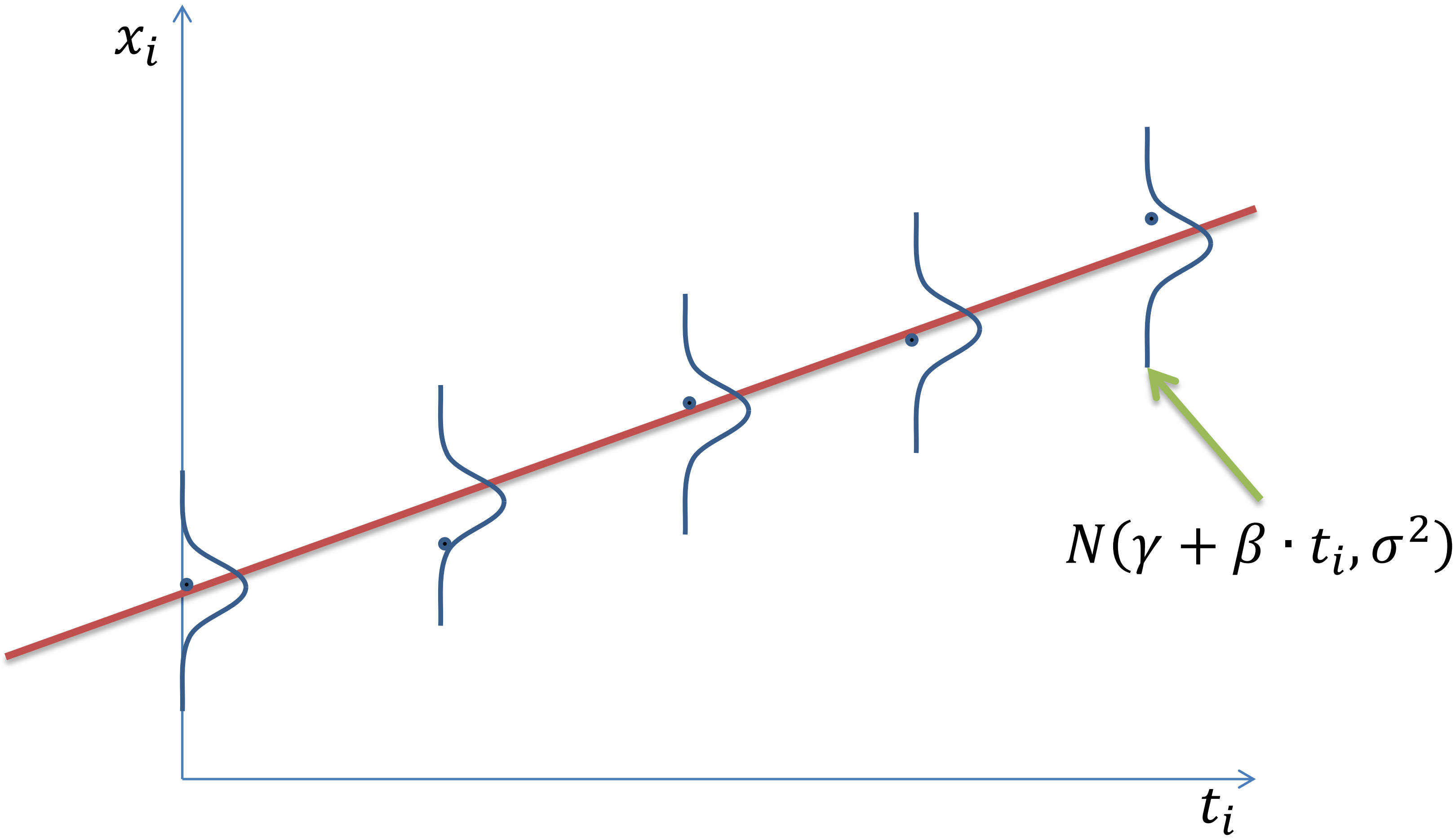
Statistisk model

- Data kommer i par (t_i, x_i), således at responserne

$$X_i \sim N(\gamma + \beta \cdot t_i, \sigma^2), i = 1, \dots, n$$

  er uafhængige stokastiske variable.
- De forklarende variable (t_i) er <u>ikke</u> stokastiske.
- Middelværdien af x_i er givet ved den lineære sammenhæng, $\gamma + \beta \cdot t_i$.
- Variansen er konstant (afhænger ikke af t_i).

Statistisk model



Parameterskøn

- Der indgår tre parametre i modellen:

$$X_i \sim N(\gamma + \beta \cdot t_i, \sigma^2), i = 1, \dots, n$$

- Hældning

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(x_i - \bar{x})}{\sum_{i=1}^{n}(t_i - \bar{t})^2}$$

- Skæring

$$\hat{\gamma} = \bar{x} - \hat{\beta} \cdot \bar{t}$$

- Empirisk varians

$$\widehat{\sigma^2} = s_r^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left[x_i - \left(\hat{\gamma} + \hat{\beta} \cdot t_i\right)\right]^2$$

Likelihood funktionen

- Modellen for målingerne

$$X_i \sim N(\gamma + \beta \cdot t_i, \sigma^2), i = 1, \dots, n$$

- giver følgende model for målefejlen:

$$\varepsilon_i = x_i - \gamma + \beta \cdot t_i \sim N(0, \sigma^2), i = 1, \dots, n$$

- Så kan tæthedsfunktionen for data, givet parametrene, skrives

$$f(x|\gamma, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \gamma - \beta \cdot t_i)^2/2\sigma^2}$$

Likelihood funktionen

- Tæthedsfunktionen kan omskrives således

$$f(x|\gamma,\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\gamma-\beta\cdot t_i)^2/2\sigma^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n}(x_i-\gamma-\beta\cdot t_i)^2/2\sigma^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-R(\gamma,\beta)/2\sigma^2}$$

hvor

$$R(\gamma,\beta) = \sum_{i=1}^{n}(x_i - \gamma - \beta \cdot t_i)^2$$

At maksimere f mht. γ og β er det samme som at minimere R.

Maximum likelihood estimater

- Differentierer vi $R(\gamma, \beta)$ med hensyn til hhv. γ og β, og sætter lig med nul, får vi maximum likelihood estimaterne:

- Hældning

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(x_i - \bar{x})}{\sum_{i=1}^{n}(t_i - \bar{t})^2}$$

- Skæring

$$\hat{\gamma} = \bar{x} - \hat{\beta} \cdot \bar{t}$$

- Man kan endvidere vise, at maximum likelihood estimatet af variansen er

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[x_i - \left(\hat{\gamma} + \hat{\beta} \cdot t_i\right)\right]^2$$

Lineær regression

- Udledning af skæringsparameteren (γ)

$$f(t_i, \gamma, \beta) = \gamma + \beta t_i$$

Find $\underset{\gamma,\beta}{\arg\min}(\varepsilon)$, where

$$\varepsilon = \sum_{i=1}^{n} \left(x_i - f(t_i, \gamma, \beta)\right)^2 = \sum_{i=1}^{n} r_i^{\,2} = \sum_{i=1}^{n} \left(x_i - \gamma - \beta t_i\right)^2$$

To find the global minimum of ε, differentiate w.r.t. γ and β and set to zero

1)

$$\frac{\partial \varepsilon}{\partial \gamma} = 2\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \gamma} = -2\sum_{i=1}^{n} \frac{\partial f(t_i, \gamma, \beta)}{\partial \gamma} r_i = -2\sum_{i=1}^{n} x_i - \gamma - \beta t_i = 0$$

$$\Updownarrow$$

$$2n\gamma = 2\sum_{i=1}^{n} x_i - 2\beta \sum_{i=1}^{n} t_i \Leftrightarrow \gamma = \frac{2}{2n}\sum_{i=1}^{n} x_i - \frac{2\beta}{2n}\sum_{i=1}^{n} t_i = \bar{x} - \beta \bar{t}$$

Hence, $\hat{\gamma} = \bar{x} - \beta \bar{t}$.

Lineær regression

- Udledning af hældningsparameteren (β)

$$\frac{\partial \varepsilon}{\partial \beta} = 2\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta} = -2\sum_{i=1}^{n} \frac{\partial f(t_i, \gamma, \beta)}{\partial \beta} r_i = -2\sum_{i=1}^{n} t_i (x_i - \gamma - \beta t_i) = -2\sum_{i=1}^{n} t_i x_i - t_i \gamma - \beta t_i^2 =$$

$$-2\sum_{i=1}^{n} t_i x_i + 2\gamma \sum_{i=1}^{n} t_i + 2\beta \sum_{i=1}^{n} t_i^2 = 0$$

Insert $\gamma = \bar{x} - \beta \bar{t}$:

$$-2\sum_{i=1}^{n} t_i x_i + 2(\bar{x} - \beta \bar{t})\sum_{i=1}^{n} t_i + 2\beta \sum_{i=1}^{n} t_i^2 = -2\sum_{i=1}^{n} t_i x_i + 2\bar{x}\sum_{i=1}^{n} t_i - 2\beta\, \bar{t}\sum_{i=1}^{n} t_i + 2\beta \sum_{i=1}^{n} t_i^2 =$$

$$-2\sum_{i=1}^{n} t_i (x_i - \bar{x}) + 2\beta \sum_{i=1}^{n} t_i (t_i - \bar{t}) = -2\sum_{i=1}^{n} (t_i - \bar{t})(x_i - \bar{x}) + 2\beta \sum_{i=1}^{n} (t_i - \bar{t})^2 = 0$$

$$\Updownarrow$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (t_i - \bar{t})(x_i - \bar{x})}{\sum_{i=1}^{n} (t_i - \bar{t})^2}$$

Testkatalog for hældningen β

- Statistisk model
  - $X_i \sim N(\gamma + \beta \cdot t_i, \sigma^2), i = 1, \dots, n$ og uafhængige.
  - Parameterskøn: $\hat{\beta} = (\sum_{i=1}^{n}(t_i - \bar{t})(x_i - \bar{x}))/(\sum_{i=1}^{n}(t_i - \bar{t})^2)$

    $\hat{\gamma} = \bar{x} - \hat{\beta} \cdot \bar{t}$

    $s_r^2 = \sum_{i=1}^{n}\left[x_i - \left(\hat{\gamma} + \hat{\beta} \cdot t_i\right)\right]^2 /(n-2)$
- Hypotesetest
  - $H: \beta = \beta_0$
  - Teststørrelse: $t = \frac{\hat{\beta} - \beta_0}{s_r \sqrt{1/\sum_{i=1}^{n}(t_i - \bar{t})^2}} \sim t(n-2)$
  - P-værdi: $pval = 2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$
- 95% konfidensinterval
  - $[\beta_-; \beta_+] = \left[\hat{\beta} - t_0 \cdot s_r \sqrt{1/\sum_{i=1}^{n}(t_i - \bar{t})^2}\,; \hat{\beta} + t_0 \cdot s_r \sqrt{1/\sum_{i=1}^{n}(t_i - \bar{t})^2}\right]$
  - Hvor $t_0 = t_{inv}(0{,}975, n-2)$

Testkatalog for skæringen γ

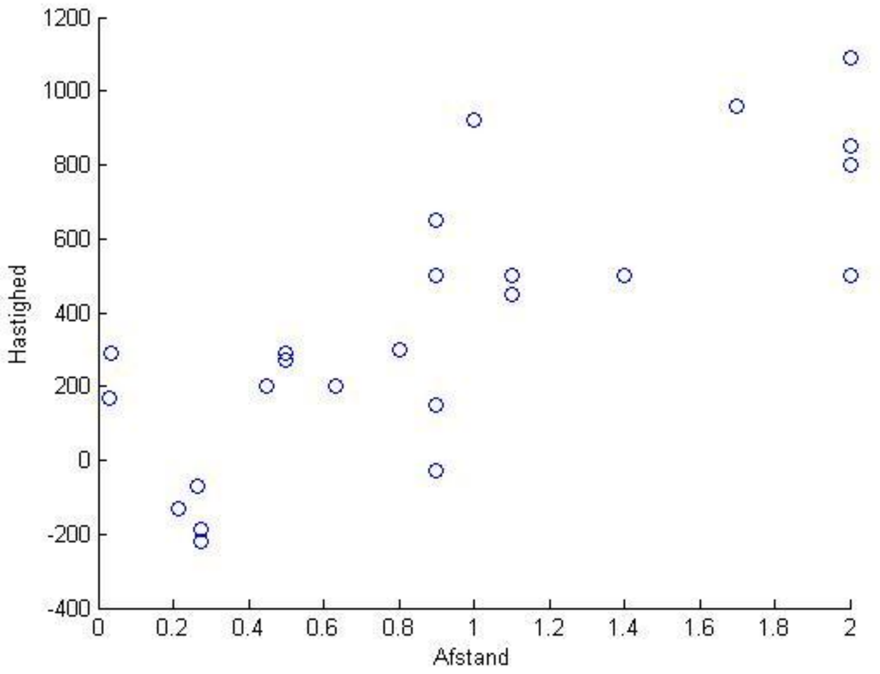
- Statistisk model
  - $X_i \sim N(\gamma + \beta \cdot t_i, \sigma^2), i = 1, \dots, n$ og uafhængige.
  - Parameterskøn: $\hat{\beta} = \left(\sum_{i=1}^{n}(t_i - \bar{t})(x_i - \bar{x})\right) / \left(\sum_{i=1}^{n}(t_i - \bar{t})^2\right)$

    $\hat{\gamma} = \bar{x} - \hat{\beta} \cdot \bar{t}$

    $s_r^2 = \sum_{i=1}^{n}\left[x_i - \left(\hat{\gamma} + \hat{\beta} \cdot t_i\right)\right]^2 / (n-2)$
- Hypotesetest
  - $H: \gamma = \gamma_0$
  - Teststørrelse: $t = \dfrac{\hat{\gamma} - \gamma_0}{s_r\sqrt{1/n + \bar{t}^2 / \sum_{i=1}^{n}(t_i - \bar{t})^2}} \sim t(n-2)$
  - P-værdi: $pval = 2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$
- 95% konfidensinterval
  - $[\gamma_-; \gamma_+] = \left[\hat{\gamma} - t_0 \cdot s_r\sqrt{\frac{1}{n} + \frac{\bar{t}^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}}; \hat{\gamma} + t_0 \cdot s_r\sqrt{\frac{1}{n} + \frac{\bar{t}^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}}\right]$
  - Hvor $t_0 = t_{inv}(0{,}975, n-2)$

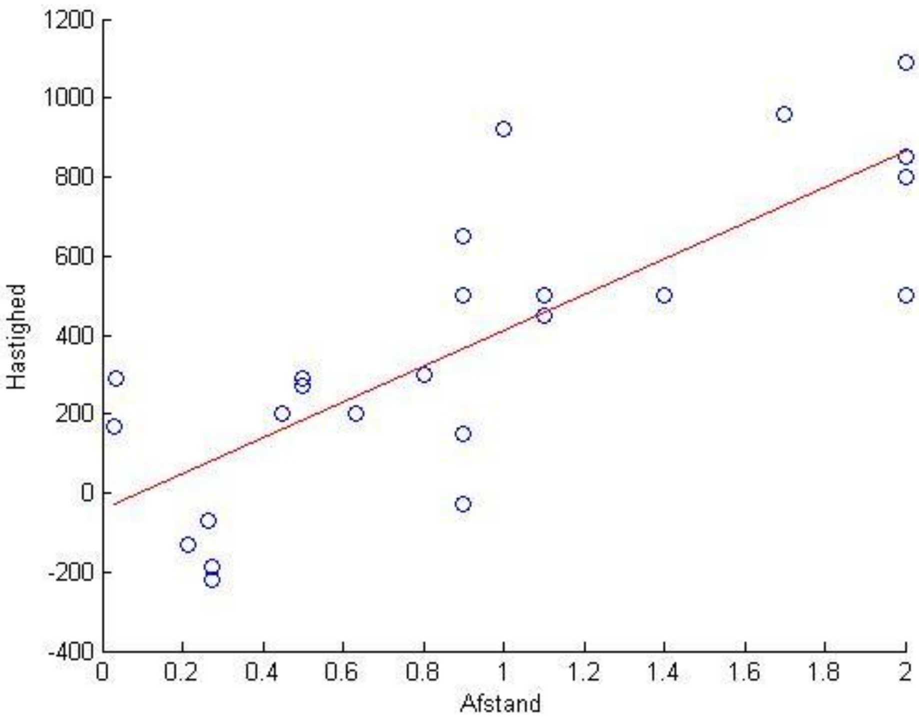
Hubbles lov

```
%% Eksempel 1 - Lineær regression - Hubble's målinger
Afstand = [ 0.032 0.034 0.214 0.263 0.275 0.275 0.450 0.500 ...
            0.500 0.630 0.800 0.900 0.900 0.900 0.900 1.000 ...
            1.100 1.100 1.400 1.700 2.000 2.000 2.000 2.000 ];

Hastighed = [ 170 290 -130 -70 -185 -220 200 290 ...
              270 200 300 -30 650 150 500 920 ...
              450 500 500 960 500 850 800 1090 ];
```



Hubbles lov

```
t = Afstand;
x = Hastighed;
n = length(t)

% Parameterskøn
beta_hat = sum((t-mean(t)).*(x-mean(x)))/sum((t-mean(t)).^2)
lambda_hat = mean(x) - beta_hat*mean(t)
```



Hubbles lov

```
% 95% konfidensinterval for hældningen
t0 = tinv(0.975,n-2)
beta_nedre = beta_hat - t0*sr*sqrt(1/sum((t-mean(t)).^2))
beta_oevre = beta_hat + t0*sr*sqrt(1/sum((t-mean(t)).^2))
```

Resultat:

```
beta_hat =
  454.1584

beta_nedre =
  298.1262

beta_oevre =
  610.1906
```

Hubbles lov

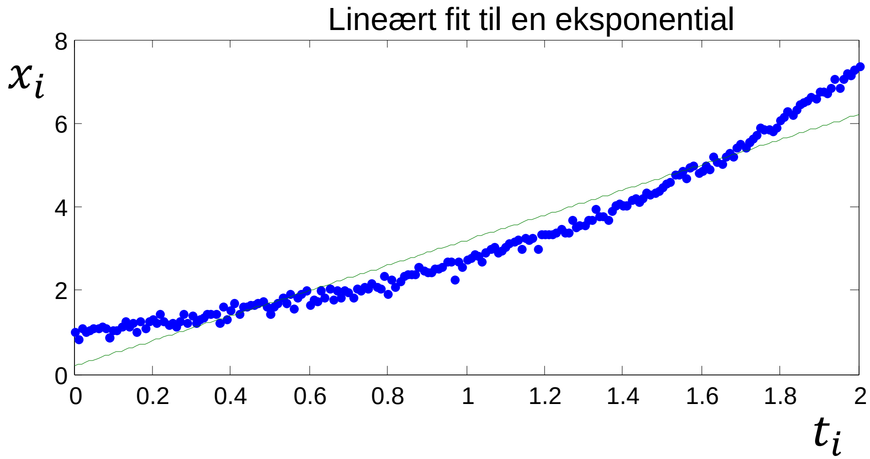
```
% 95% konfidensinterval for skæringen
t0 = tinv(0.975,n-2)
lambda_nedre = lambda_hat - t0*sr*sqrt(1/n+mean(t)^2/sum((t-mean(t)).^2))
lambda_oevre = lambda_hat + t0*sr*sqrt(1/n+mean(t)^2/sum((t-mean(t)).^2))
```

Resultat:

```
lambda_hat =
  -40.7836

lambda_nedre =
 -213.8253

lambda_oevre =
  132.2580
```

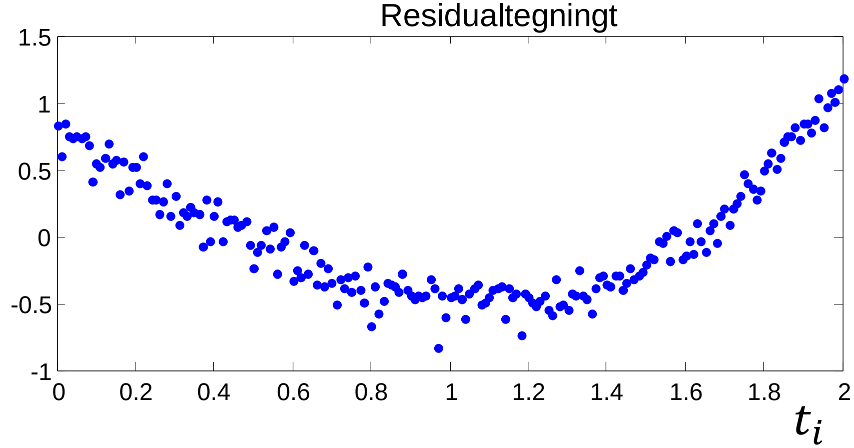
Pas på!

- Lineært fit til en eksponentialfunktion



Residual:

$$r_i = x_i - (\hat{\gamma} + \hat{\beta} \cdot t_i)$$



Residualerne er <u>ikke</u> tilfældige - de afhænger af t!

Residualtegning

- Plot

$$t_i \text{ vs. } r_i$$

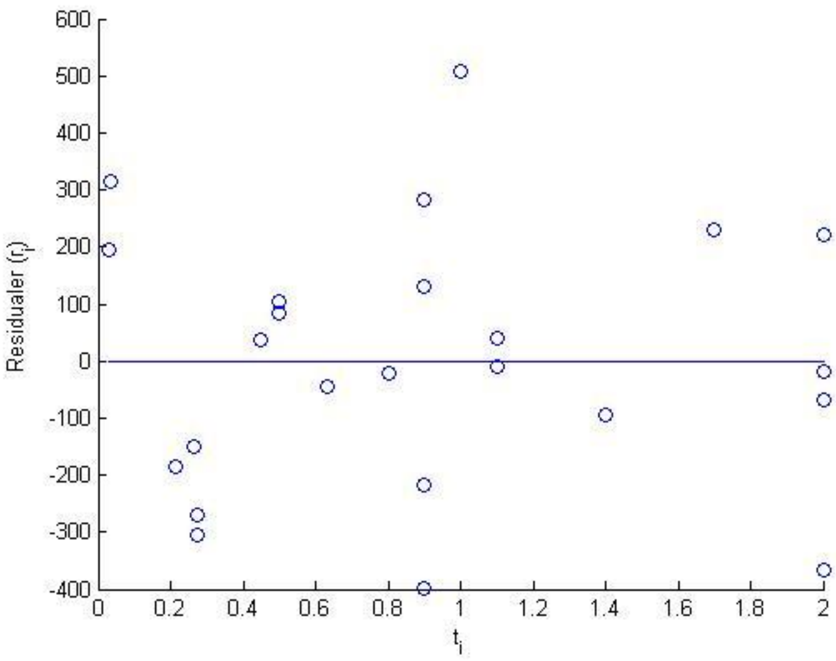
- Hvor

$$r_i = x_i - \left(\hat{\gamma} + \hat{\beta} \cdot t_i\right)$$

- Vi kigger efter to ting:
  - Værdierne af residualerne $(x_i - \widehat{x_i})$ må ikke afhænge af t_i, men skal ligge tilfældigt omkring 0.
  - Variansen af residualerne må ikke afhænge af t_i.

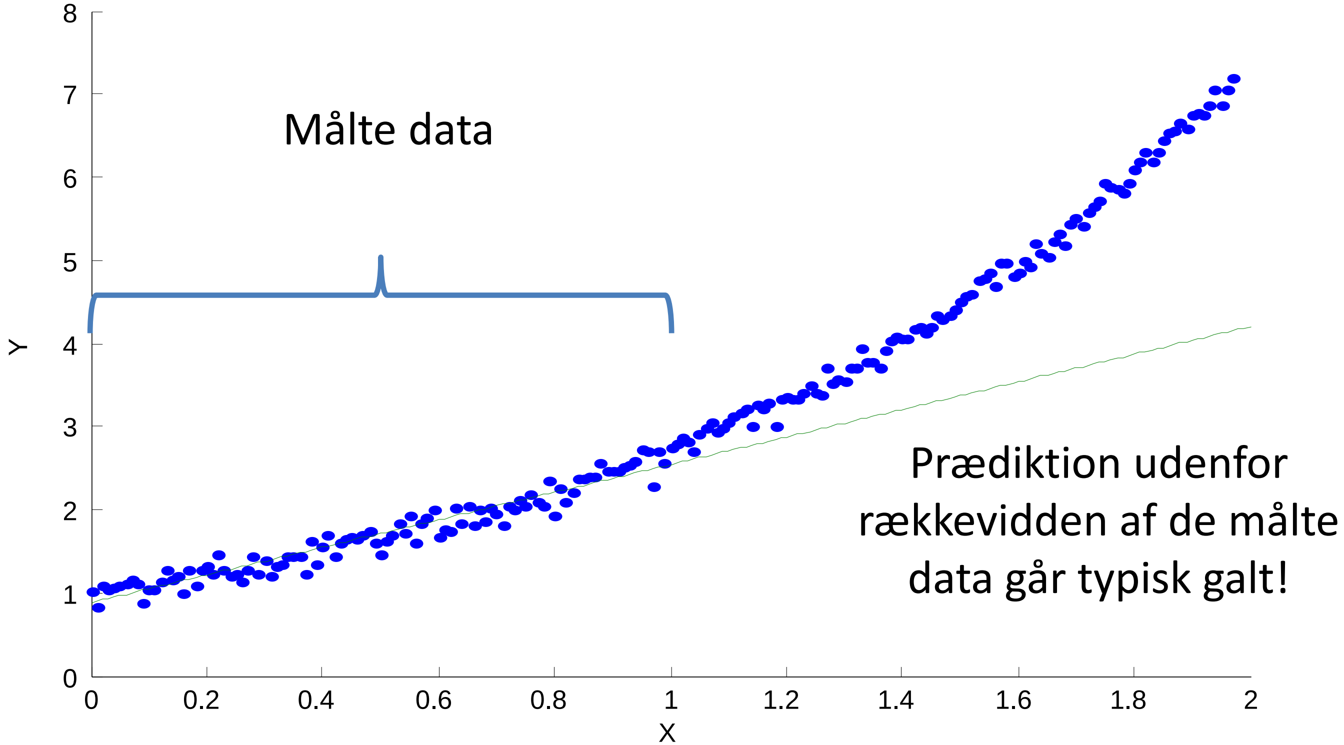
Residualtegning for Hubbles eksperiment

```
% Residualtegning
figure
scatter(t,x-(lambda_hat+beta_hat*t))
hold on
plot([t_min t_max],zeros(1,2))
hold off
xlabel('t_i')
ylabel('Residualer (r_i)')
```



Brug af lineær regression til prædiktion

- Vi kan bruge den lineære model til at prædiktere en *x*-værdi, givet *t*.
- MEN: Pas på med at prædiktere ud over det interval, du har brugt til at fitte modellen:



Correlation

- How do we estimate the strength of a linear relation?
- The correlation coefficient:

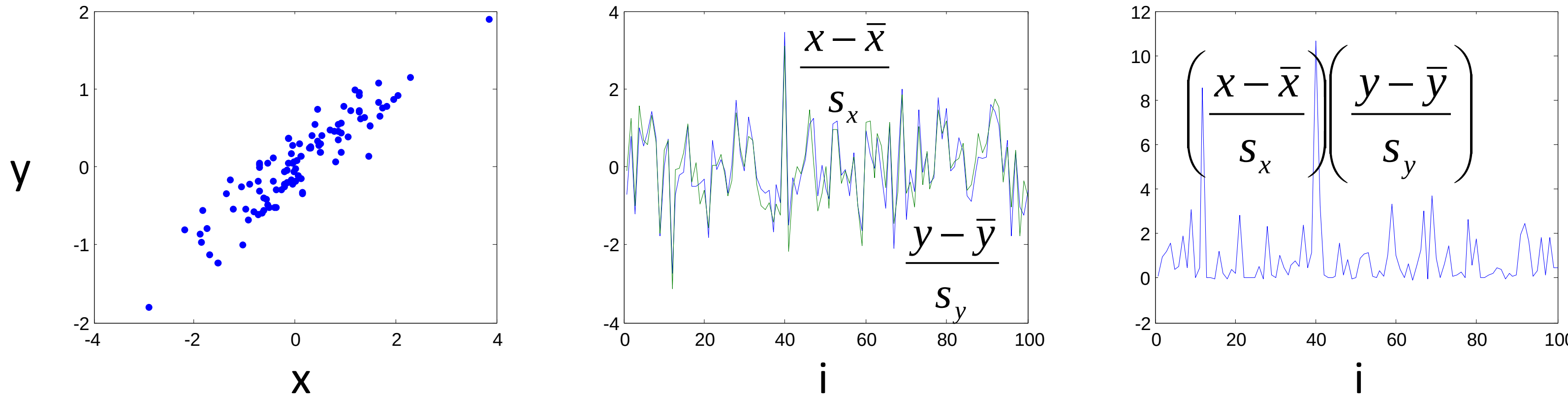
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) = \frac{1}{n-1}\sum_{i=1}^{n} z_x z_y$$

z-scores

  - r takes on values from -1 to 1
  - Perfect positive linear correlation, r=1
  - Perfect negative linear correlation, r=-1
  - No correlation, r=0

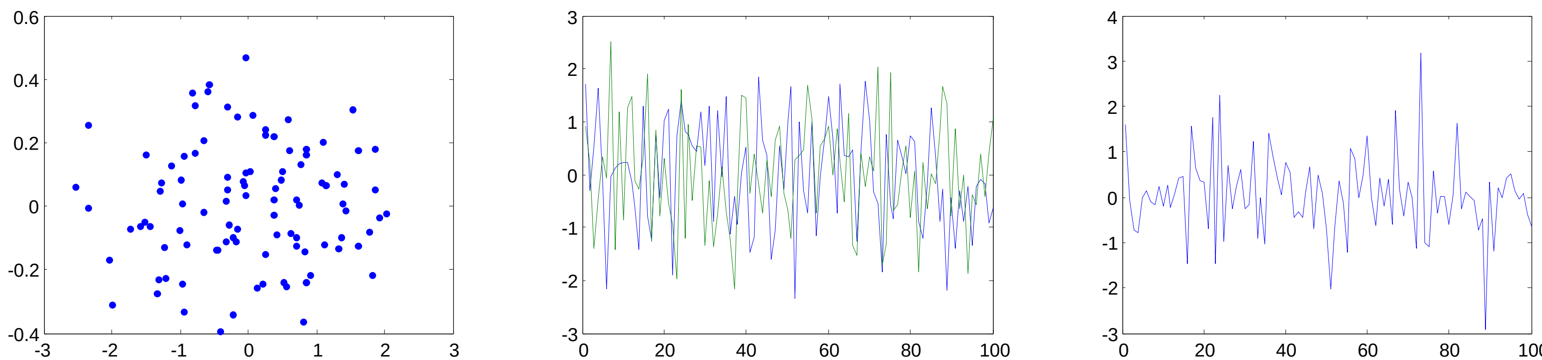
Correlation

Strong correlation



$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Weak correlation



- Recall that

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) = \frac{1}{n-1}\sum_{i=1}^{n} z_x z_y$$

x	**y**	$\mathbf{z_x}$	$\mathbf{z_y}$	$\mathbf{z_x z_y}$
0.9490	0.6982	1.1671	1.5141	1.7671
0.1996	0.3499	0.2506	0.7715	0.1934
0.4997	0.3819	0.6176	0.8397	0.5187
-0.7936	-0.4946	-0.9639	-1.0292	0.9920
-0.7832	-0.3331	-0.9513	-0.6848	0.6514
-1.5193	-0.4773	-1.8515	-0.9921	1.8368
0.3533	0.1733	0.4386	0.3950	0.1732
-0.0291	-0.3031	-0.0291	-0.6207	0.0180
-0.4422	-0.2400	-0.5343	-0.4862	0.2598
0.0653	-0.1616	0.0864	-0.3190	-0.0276
-0.7394	-0.2227	-0.8977	-0.4494	0.4034

The large $\mathbf{z_x z_y}$ score indicates that these points contribute a considerable amount to the correlation coefficient

Correlation

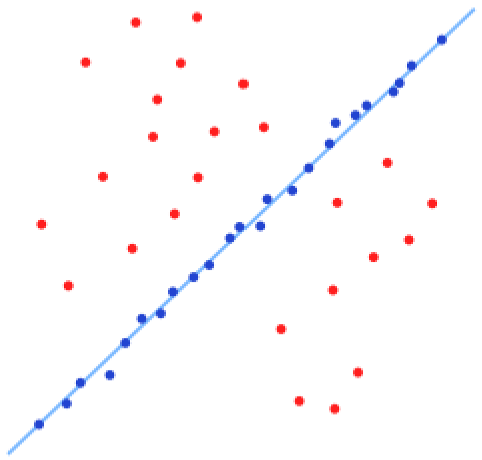
- Watch out!
  - Points with a high $\mathbf{z_x z_y}$ score are separated from the rest of the data and are *potentially influential* (i.e., outliers).
  - Outliers can have a dramatic effect on the correlation coefficient.
  - The extent of influence of any point can be judged in part, by computing the correlation coefficient with and without that point.

Outliers

- Outliers
  - Data points that have a very strong influence.
  - Can mask linear relationships
  - … or give a false impression of a linear relationship
- How to handle outliers?
  - Exclude them from the analysis
  - In regression analysis, use a ”robust” technique that is less sensitive to outliers
  - "RANdom SAmple Consensus“ (RANSAC) achieves this by iteratively working on a random subset of the original data.

RANSAC

A data set with many outliers for which a line has to be fitted.



Fitted line with RANSAC*, outliers have no influence on the result.

* Source: http://en.wikipedia.org/wiki/RANSAC

RANSAC

- RANSAC* achieves its goal by iteratively selecting a random subset of the original data. These data are *hypothetical inliers* and this hypothesis is then tested as follows:
  - A model is fitted to the hypothetical inliers, i.e. all free parameters of the model are reconstructed from the inliers.
  - All other data are then tested against the fitted model and, if a point fits well to the estimated model, also considered as a hypothetical inlier.
  - The estimated model is reasonably good if sufficiently many points have been classified as hypothetical inliers.
  - The model is re-estimated from all hypothetical inliers, because it has only been estimated from the initial set of hypothetical inliers.
  - Finally, the model is evaluated by estimating the error of the inliers relative to the model.

Correlation

- Using the correlation coefficient to explore relationships
- Ex. 2.10: pollutant data of 15 US cities in year 2000

	CO	O_3	PM_{10}	SO_2
CO	1	0.87	0.36	0.17
O_3	0.87	1	0.20	0.098
PM_{10}	0.36	0.20	1	0.091
SO_2	0.17	0.098	0.091	1

Correlation coefficients

  - All four pollutants are positively correlated
  - However, the levels of CO are most strongly correlated with O_3 levels
- A word of caution
  - Do not infer a causal relationship on the basis of high correlation!

Functional MRI (fMRI)

- MRI – Magnetic Resonance Imaging

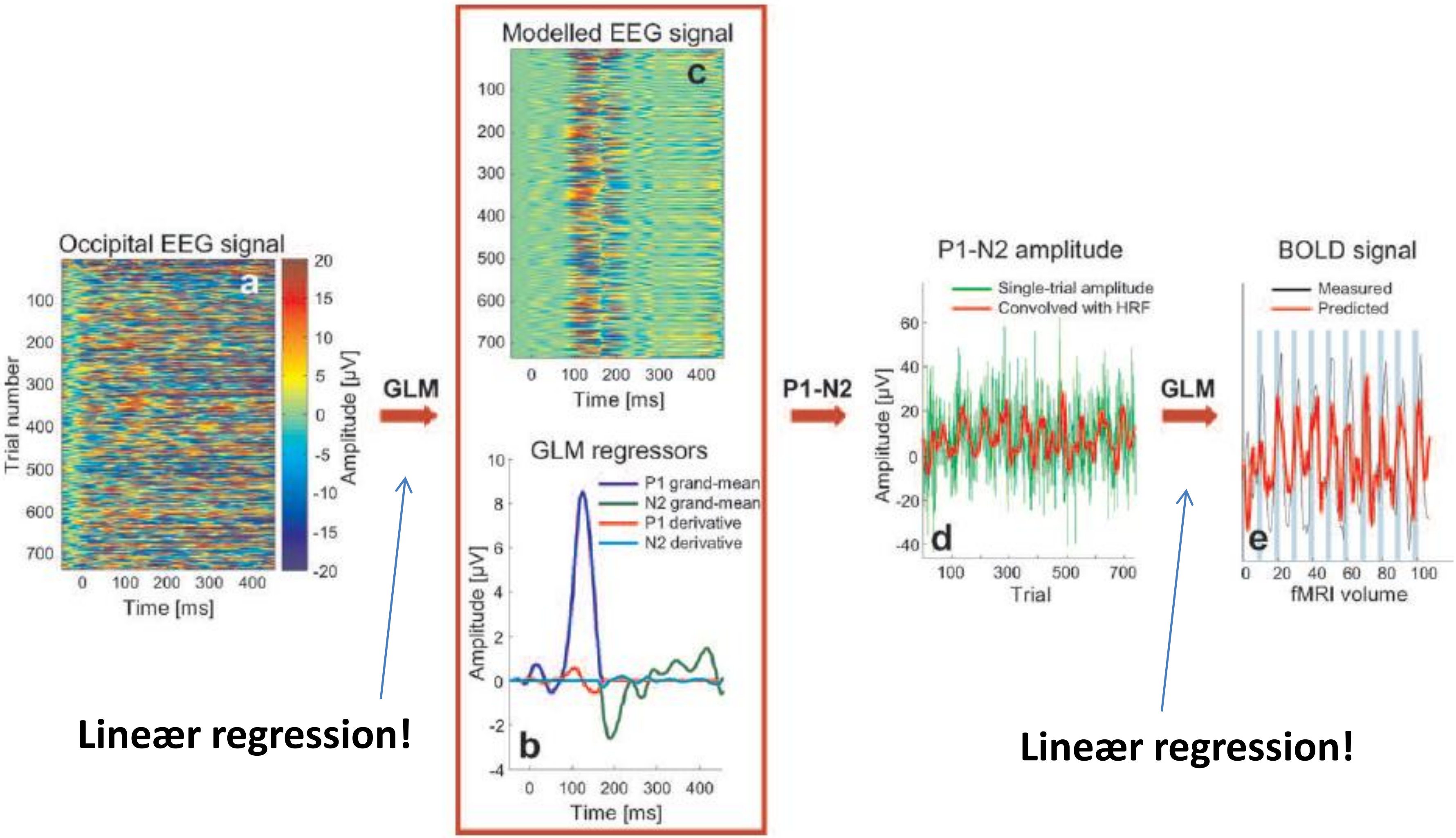
Visual fMRI

Magnetic Resonance in Medicine 68:252–260 (2012)

Correlation Between Single-Trial Visual Evoked Potentials and the Blood Oxygenation Level Dependent Response in Simultaneously Recorded Electroencephalography–Functional Magnetic Resonance Imaging

Dan Fuglø,[1*] Henrik Pedersen,[2] Egill Rostrup,[1] Adam E. Hansen,[1,3] and Henrik B. W. Larsson[1,4,5]

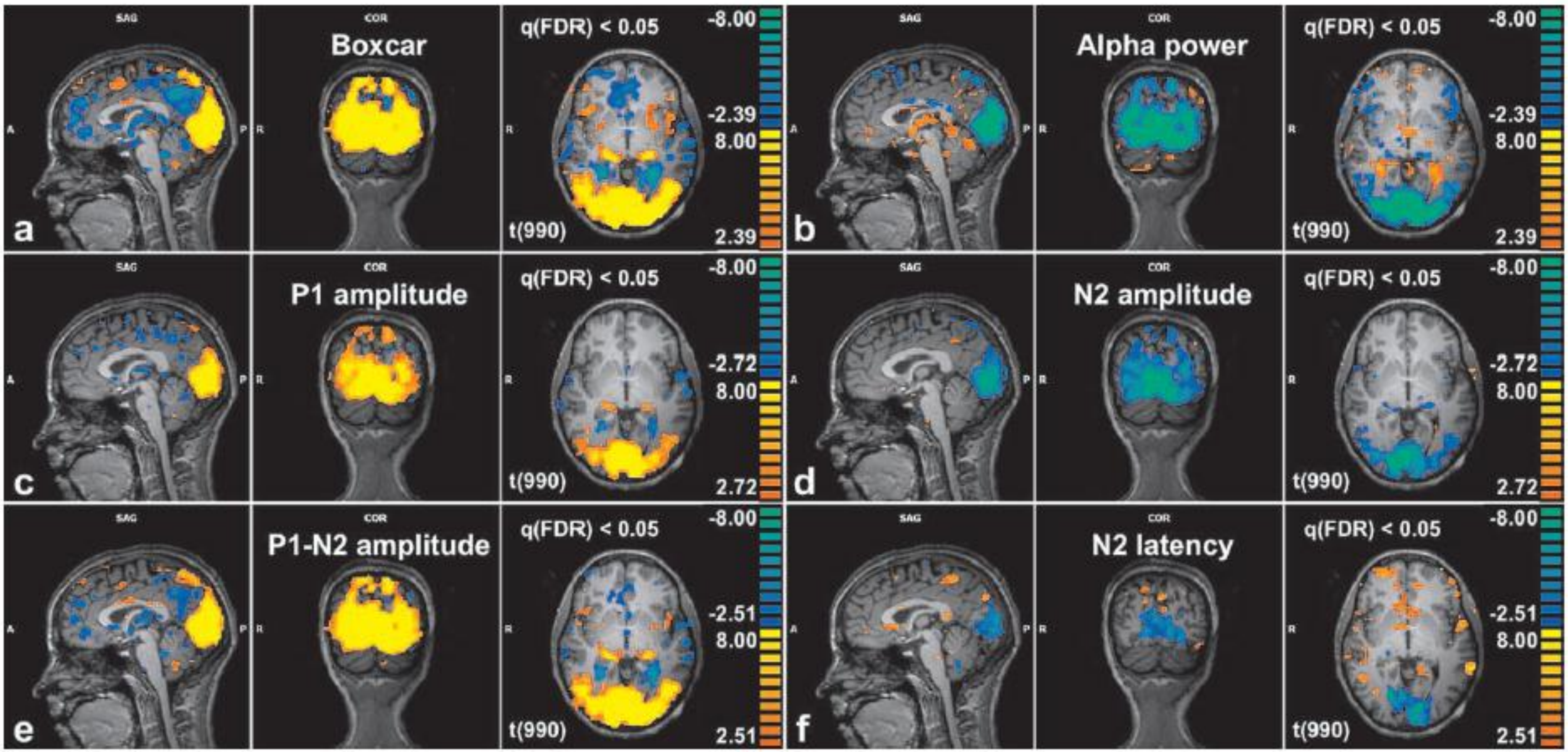
Visual fMRI



Lineær regression!

Lineær regression!

Visual fMRI



Boxcar
q(FDR) < 0.05
t(990)
a
Alpha power
b
P1 amplitude
c
N2 amplitude
d
P1-N2 amplitude
e
N2 latency
f
-8.00
-2.39
8.00
2.39
-2.72
2.72
-2.51
2.51