UNIVERSITEIT GENT

BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

H6: Enkelvoudige lineaire regressie

Lieven Clement

Statistiek: 2de bach. in de Biochemie en Biotechnologie, Biologie, Biomedische Wetenschappen, en de Chemie

Borstkanker dataset (subset van studie https://doi.org/10.1093/jnci/djj052)

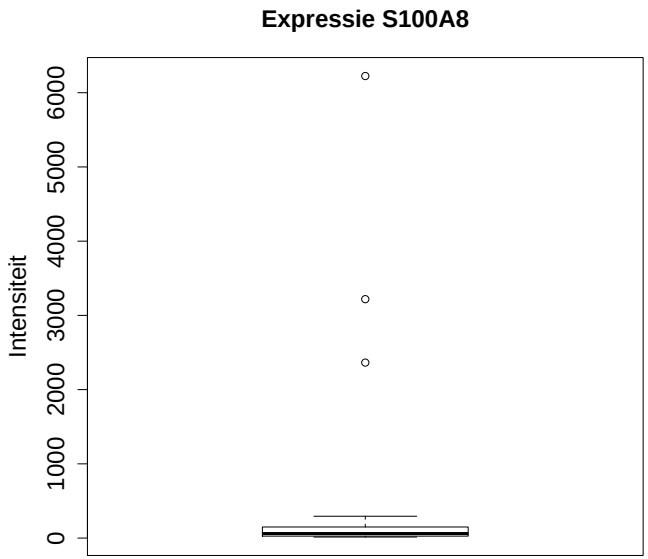
```
> borstkanker=read.table("borstkanker.txt",header=TRUE)
> head(borstkanker)
  sample_name        filename treatment er grade node size age      ESR1    S100A8
1    OXFT_209 gsm65344.cel.gz tamoxifen  1     3    1  2.5  66 1939.1990 207.19682
2   OXFT_1769 gsm65345.cel.gz tamoxifen  1     1    1  3.5  86 2751.9521 36.98611
3   OXFT_2093 gsm65347.cel.gz tamoxifen  1     1    1  2.2  74  379.1951 2364.18306
4   OXFT_1770 gsm65348.cel.gz tamoxifen  1     1    1  1.7  69 2531.7473 23.61504
...
```

- 32 borstkanker patiënten met een estrogen recepter positieve tumor (response op hormonen) die tamoxifen chemotherapy behandeling ondergaan.
- Variabelen:
  - grade: histologische graad van tumor (graad 1 vs 3),
  - node: status van de lymfe knopen (0: niet aangetast, 1: aantasting en verwijdering van de lymfe knopen),
  - size: grootte van tumor in cm,
  - ESR1 en S100A8 gen expressie in tumor biopsy (via microarray technologie)

Borstkanker dataset

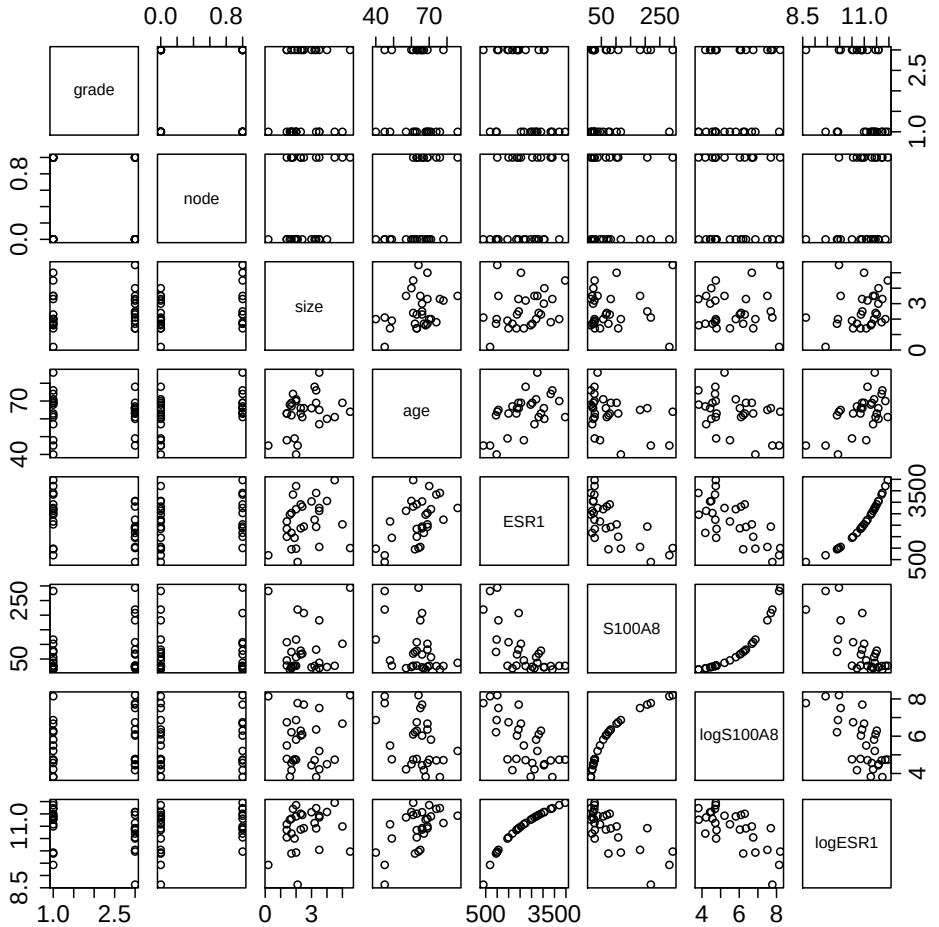
Omwille van didactische redenen zullen we eerst 3 outliers in de S100A8 expressie data verwijderen. In deze studie kan dit echter niet worden verantwoord. Later in de les zullen we aangeven hoe correct met alle data kan worden omgegaan.

```
> boxplot(borstkanker$S100A8,ylab="Intensiteit",main="Expressie S100A8")
```



Grafische voorstelling van meerdere numerieke variabelen

```
> plot(subset(borstkanker,S100A8<2000)[,-(1:4)])
```

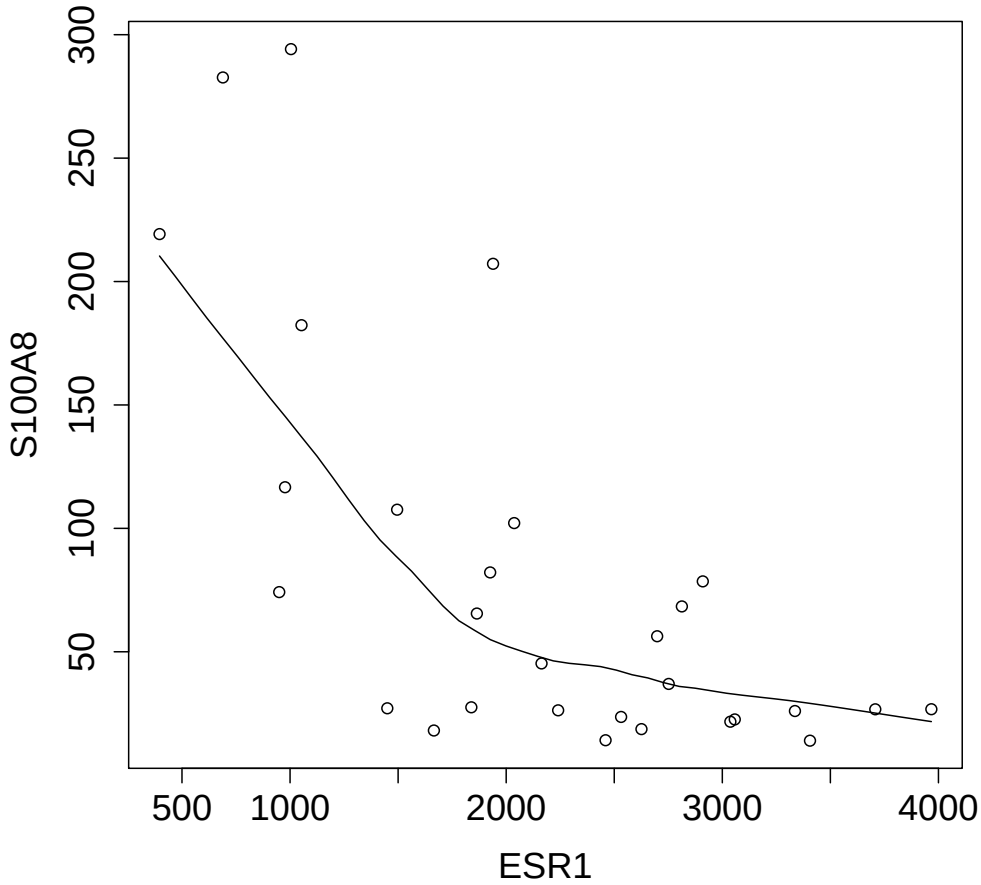


Associatie tussen ESR1 en S100A8 expressie

- ESR1 komt in $\pm$ 75% van de borstkankertumoren tot expressie. Expressie van het oestrogeen recepter gen is positief voor de behandeling omdat de kanker dan vatbaar is voor hormoontherapie. Tamoxifen gaat bijvoorbeeld interageren met de de ER recepter en genexpressie moduleren.
- Proteinen van de S100 familie zijn vaak gedisreguleerd in cancer. Expressie van S100A8 in tumor weefsel is ondermeer betrokken in het onderdrukken van het immuunsysteem in de tumor en het creëren van een inflamatoir milieu die kankergroei promoot.
- Interesse in associatie tussen ESR1 en S100A8 expressie.
- Regressiecurve, bvb. scatterplot smoother.

Scatterplot smoother

```
> with(subset(borstkanker,S100A8<2000), scatter.smooth(ESR1,S100A8))
```



Pearson correlatie

Pearson correlatie

drukt associatie tussen continue variabelen uit:

$$\text{Cor}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

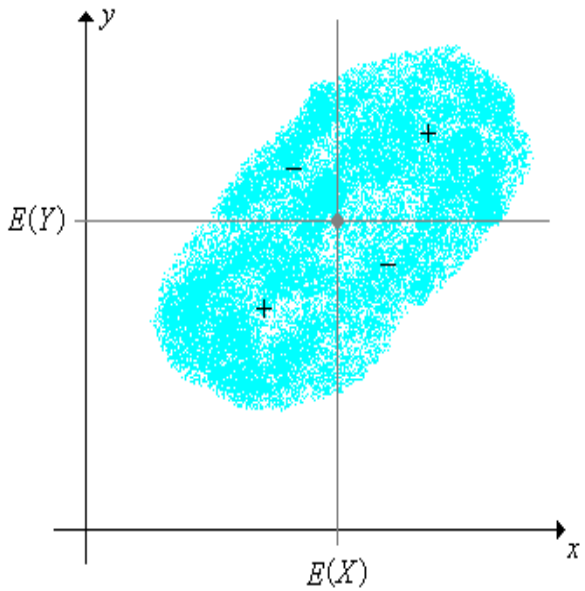
Pearson correlatie

Pearson correlatie

drukt associatie tussen continue variabelen uit:

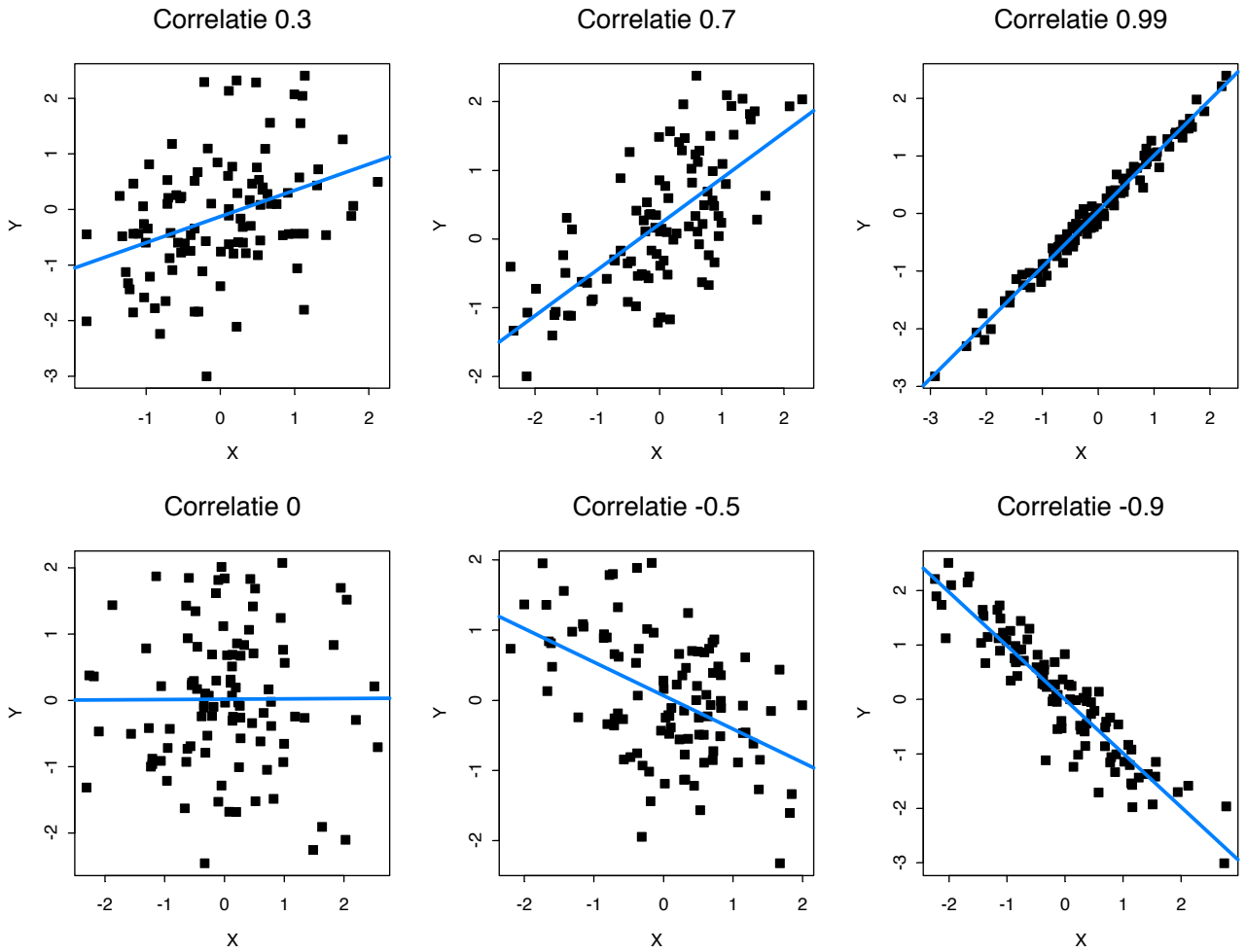
$$\text{Cor}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- Positieve correlatie: $x \nearrow \Rightarrow y \nearrow$
- Negatieve correlatie: $x \nearrow \Rightarrow y \searrow$
- Correlatie ligt steeds tussen -1 en 1.

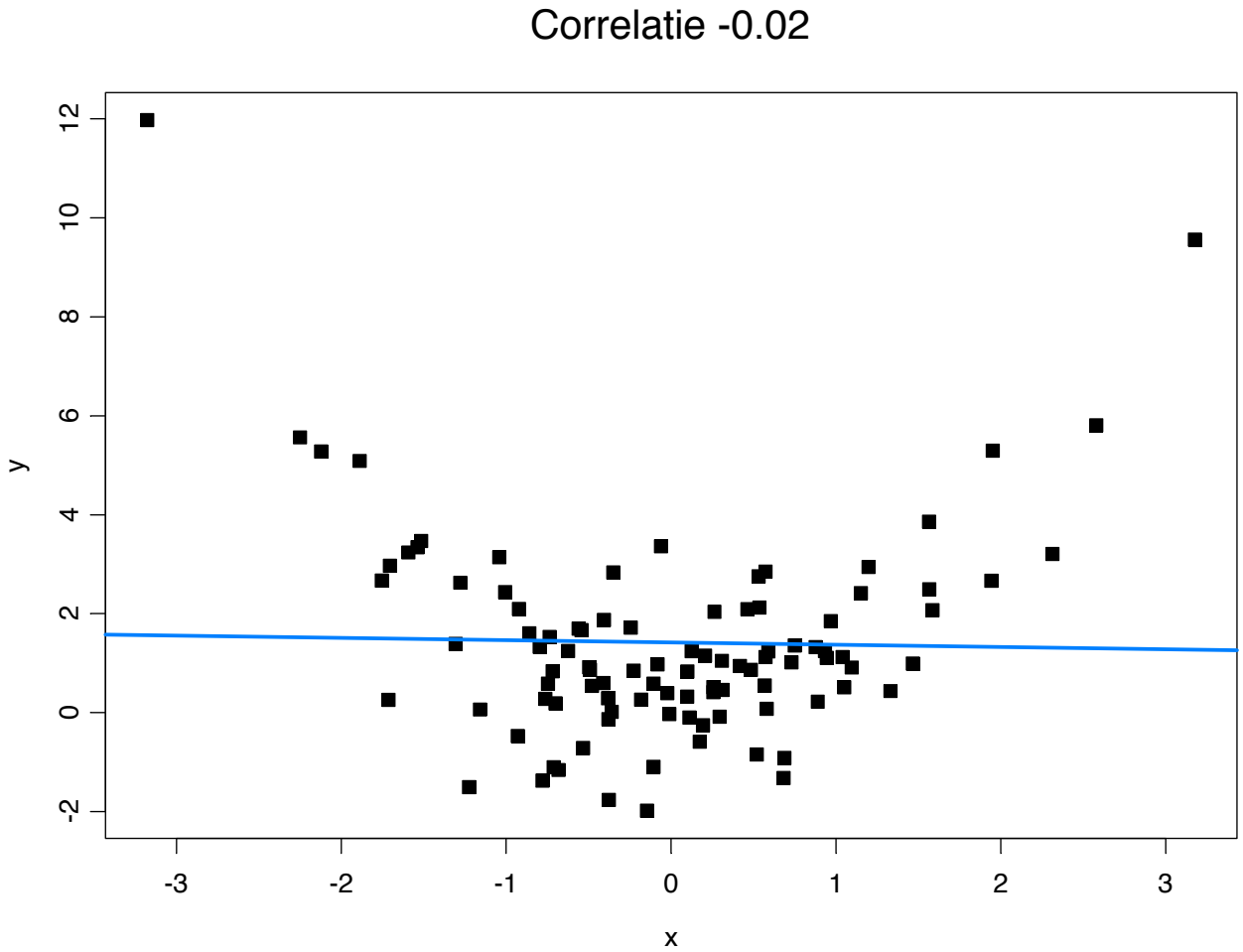


Correlatie tussen ESR1 en S100A8 expressie: -0.69.

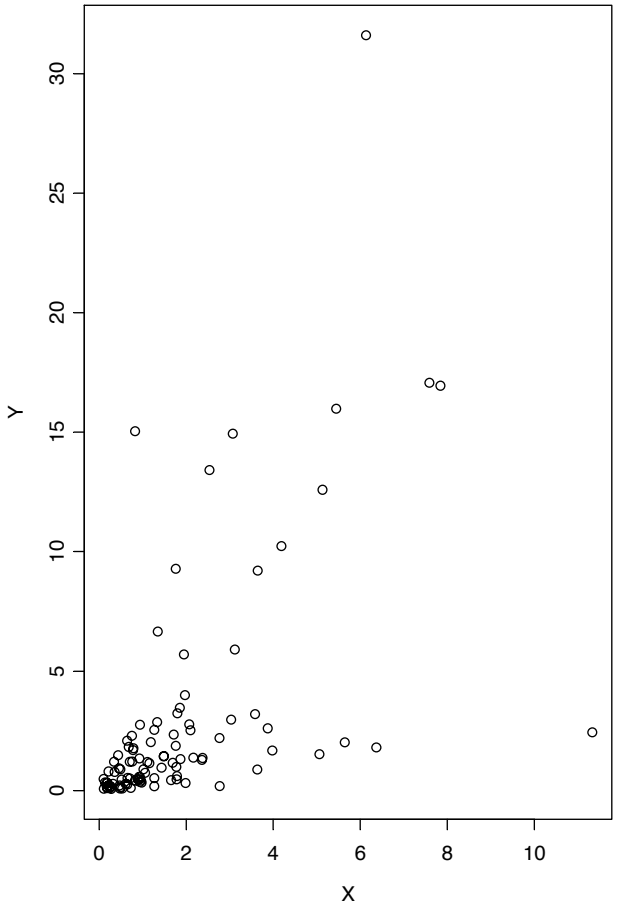
```
> with(subset(borstkanker,S100A8<2000),cor(ESR1,S100A8))
[1] -0.6854281
```

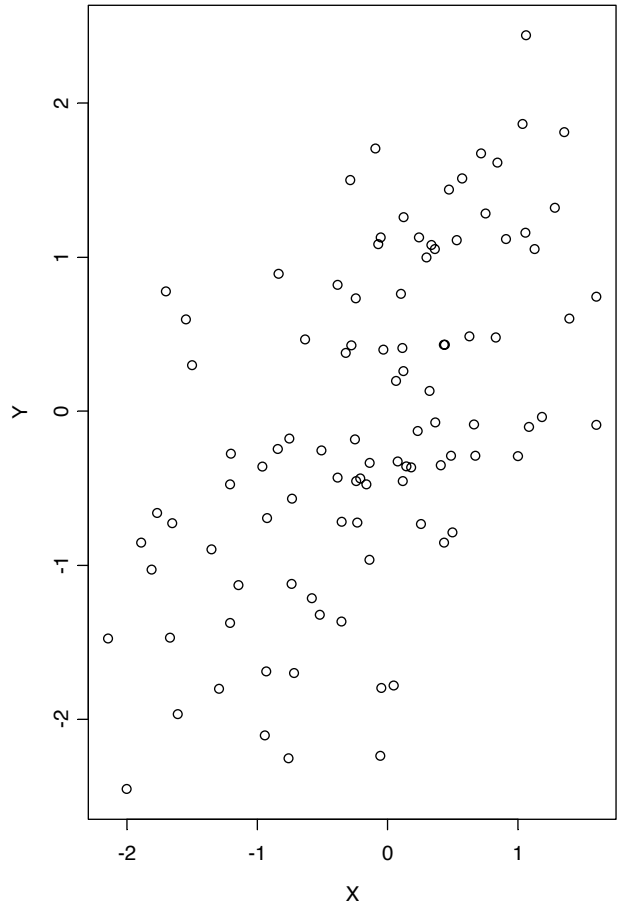


Correlatie 0 betekent 'geen lineaire associatie'

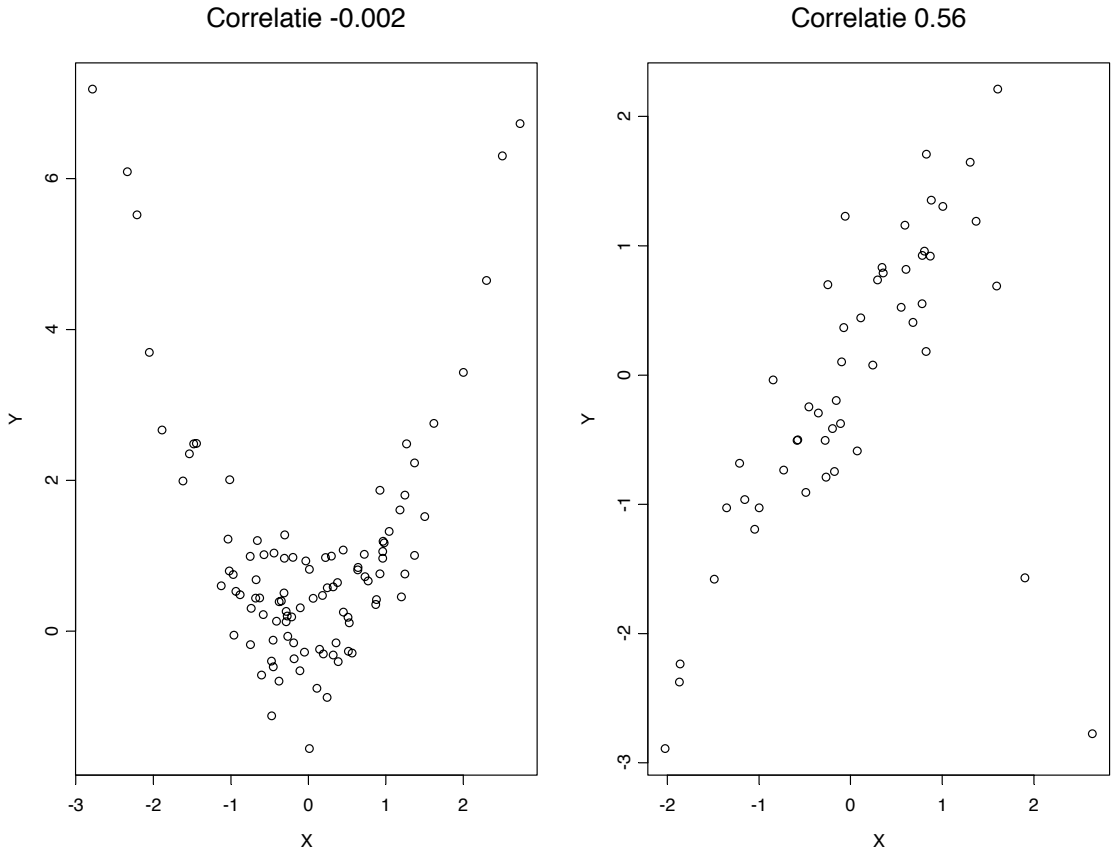


Correlatie voor scheve verdelingen





Correlatie gevoelig aan outliers




```
> with(borstkanker,cor(ESR1,S100A8))
[1] -0.5376479
>> with(borstkanker,cor(ESR1,S100A8,method="spearman"))
[1] -0.733871
```

Enkelvoudige lineaire regressie

Regressie (1)

Statistische methode met als doel de relatie tussen 2 reeksen observaties (X_i, Y_i), bekomen voor onafhankelijke subjecten $i = 1, ..., n$, te beschrijven.

Regressie (1)

Statistische methode met als doel de relatie tussen 2 reeksen observaties (X_i, Y_i), bekomen voor onafhankelijke subjecten $i = 1, ..., n$, te beschrijven.

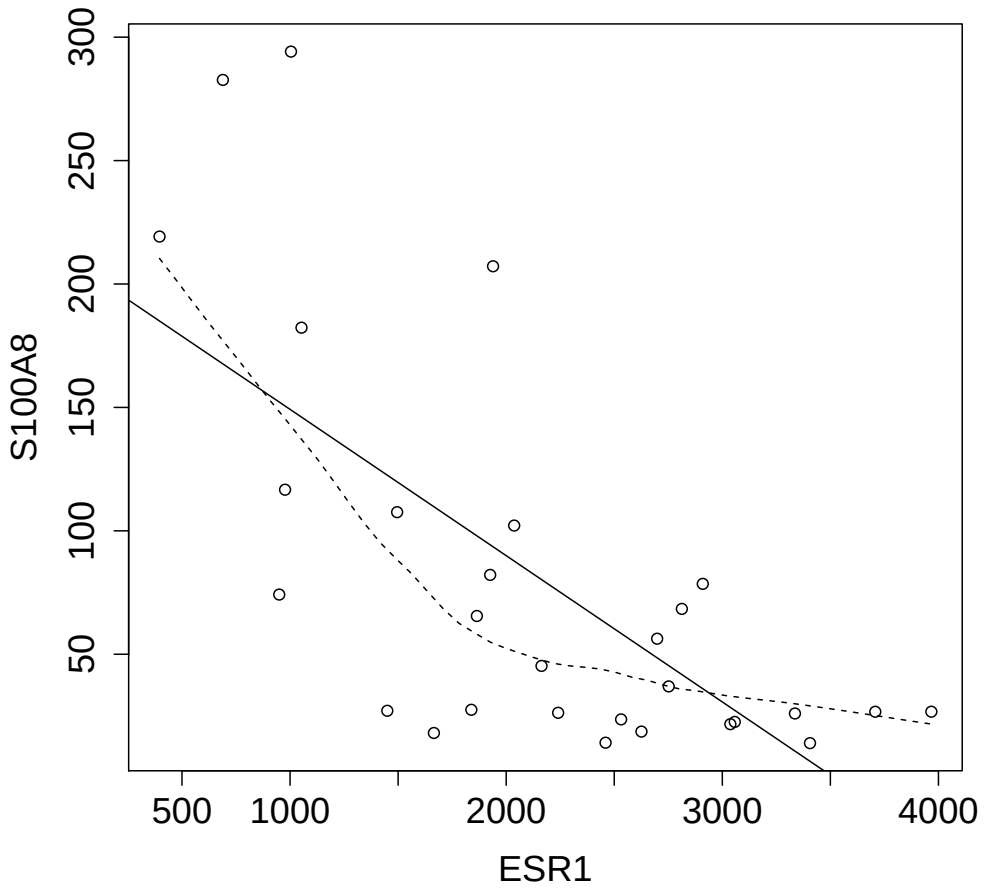
voorbeeld

Gen-expressie.

- Afhankelijke variabele, uitkomst, respons Y: S100A8 expressie.
- Onafhankelijke variabele, verklarende variabele, predictor X: ESR1 expressie.

Scatterplot met lokale regressie smoother

```
> with(subset(borstkanker,S100A8<2000), scatter.smooth(ESR1,S100A8,lpars=list(lty=2)))
> abline(lm(S100A8~ESR1,data=subset(borstkanker,S100A8<2000)))
```



Regressie (2)

- Bij vaste X neemt Y niet noodzakelijk steeds dezelfde waarde aan:

$$\text{observatie} = \text{signaal} + \text{ruis}$$

Regressie (2)

- Bij vaste X neemt Y niet noodzakelijk steeds dezelfde waarde aan:

$$\text{observatie} = \text{signaal} + \text{ruis}$$

- Observaties wiskundig modelleren als

$$Y_i = g(X_i) + \epsilon_i$$

- We definiëren $g(x)$ als de verwachte uitkomst bij subjecten met $X_i = x$

$$E(Y_i|X_i = x) = g(x)$$

zodat ϵ_i gemiddeld 0 is bij subjecten met dezelfde X_i:

$$E(\epsilon_i|X_i) = 0$$

Lineaire regressie

- Om accurate en interpreteerbare resultaten te bekomen, kiest men $g(x)$ vaak als een lineaire functie van ongekende parameters.
- Men werkt dan met lineair regressiemodel

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

voor onbekend intercept β_0 en helling β_1.

Lineaire regressie

- Om accurate en interpreteerbare resultaten te bekomen, kiest men $g(x)$ vaak als een lineaire functie van ongekende parameters.
- Men werkt dan met lineair regressiemodel

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

voor onbekend intercept β_0 en helling β_1.

- Lineair regressiemodel legt onderstelling op de verdeling van X en Y, en kan bijgevolg vals zijn.
- Als de onderstelling opgaat, laat ze efficiënte data-analyse toe: alle observaties benut om iets te leren over verwachte uitkomst bij $X = x$.

Gebruik

- Predictie: wanneer Y ongekend is, maar X wel, kunnen we Y voorspellen op basis van X

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

  - intercept: $E(Y|X = 0) = \beta_0$

Gebruik

- Predictie: wanneer Y ongekend is, maar X wel, kunnen we Y voorspellen op basis van X

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

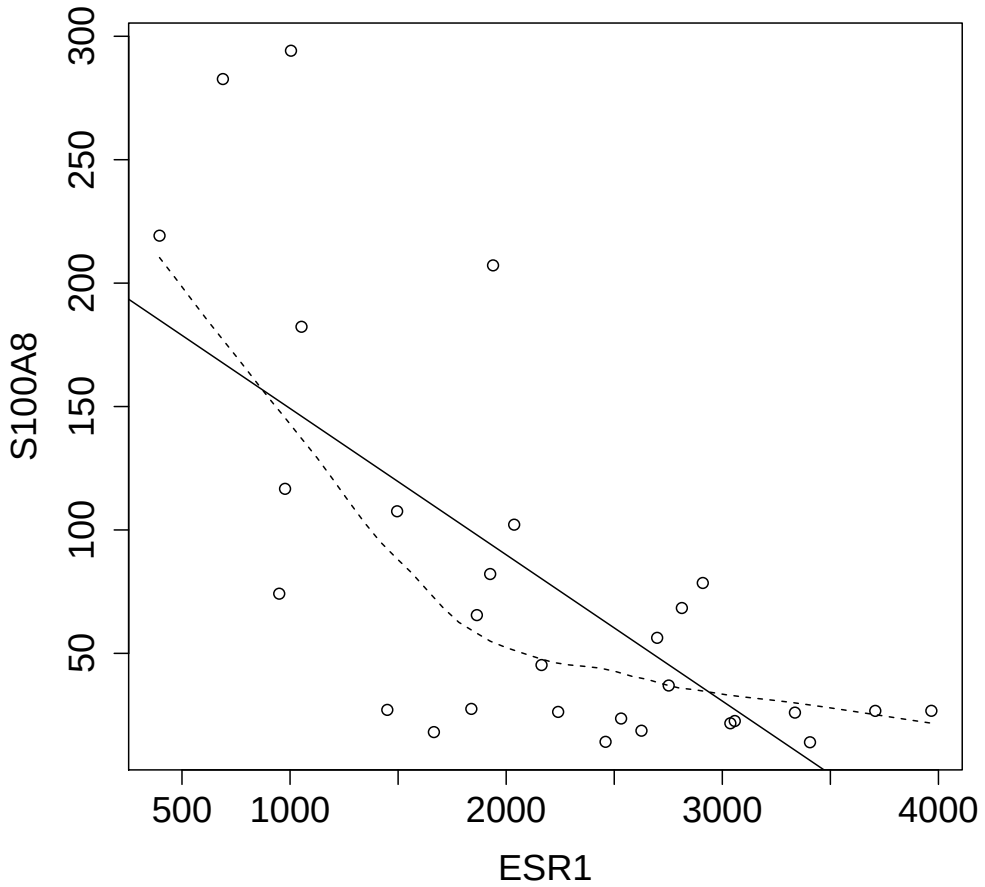
  - intercept: $E(Y|X = 0) = \beta_0$
- Associatie: biologische relatie tussen variabele X en continue meting Y beschrijven.
  - helling:

$$\begin{aligned} E(Y|X = x + \delta) - E(Y|X = x) &= \beta_0 + \beta_1(x + \delta) - \beta_0 - \beta_1 x \\ &= \beta_1 \delta \end{aligned}$$

$\beta_1 =$ verschil in gemiddelde uitkomst tussen subjecten die 1 eenheid verschillen in de waarde van X.

Parameterschatters

Parameter schatters: kleinste kwadratenschatters



Kleinste kwadratenschatters

- Kleinste kwadratenlijn: de lijn die 'het best past' bij de gegevens.
- Deze vindt men door door deze waarden voor β_0 en β_1 te kiezen die de afstand

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} e_i^2$$

  zo klein mogelijk maakt.
- Men vindt volgende schattingen voor β_1 en β_0:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{Cor}(X, Y) S_Y}{S_X}
\end{aligned}$$

Lineaire regressie output (1)

```
> model <- lm(S100A8 ESR1,subset(borstkanker,S100A8<2000))
> summary(model)
```

```
> summary(model)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 208.47145   28.57207   7.296 7.56e-08 ***
ESR1         -0.05926    0.01212  -4.891 4.08e-05 ***
```

$$E(Y|X = x) = 208.5 - 0.056x$$

Lineaire regressie output (2)

- model

$$E(Y|X = x) = 208.5 - 0.056x$$

- De verwachte S100A8 expressie is gemiddeld 56 eenheden lager bij patiënten met een ESR1 expressieniveau die 1000 eenheden hoger ligt.

Lineaire regressie output (2)

- model

$$E(Y|X = x) = 208.5 - 0.056x$$

- De verwachte S100A8 expressie is gemiddeld 56 eenheden lager bij patiënten met een ESR1 expressieniveau die 1000 eenheden hoger ligt.
- De verwachte S100A8 intensiteit voor patiënten met een ESR1 expressie-niveau van 2000 bedraagt

$$208.5 - 0.056 \times 2000 = 96.5$$

Lineaire regressie output (2)

- model

$$E(Y|X = x) = 208.5 - 0.056x$$

- De verwachte S100A8 expressie is gemiddeld 56 eenheden lager bij patiënten met een ESR1 expressieniveau die 1000 eenheden hoger ligt.
- De verwachte S100A8 intensiteit voor patiënten met een ESR1 expressie-niveau van 2000 bedraagt

$$208.5 - 0.056 \times 2000 = 96.5$$

- De verwachte S100A8 intensiteit voor patiënten met een ESR1 expressie-niveau van 4000 is

$$208.5 - 0.056 \times 4000 = -15.5$$

- Let op voor extrapolatie! (onderstelling van lineariteit kan men enkel nagaan binnen het bereik van de data).

Besluitvorming

Besluitvorming voor eenvoudige lineaire regressie

Om besluiten te kunnen trekken over lineaire regressiemodel

$$E(Y|X) = \beta_0 + \beta_1 X$$

extra onderstellingen nodig:

- homoscedasticiteit: bij vaste X heeft Y constante variantie

$$\mathrm{Var}(Y|X) = \mathrm{E}\left[\{Y - g(x)\}^2\right] = \sigma^2$$

die we schatten als

$$MSE = \frac{\sum_{i=1}^{n} \{y_i - \hat{g}(x_i)\}^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Besluitvorming voor eenvoudige lineaire regressie

Om besluiten te kunnen trekken over lineaire regressiemodel

$$E(Y|X) = \beta_0 + \beta_1 X$$

extra onderstellingen nodig:

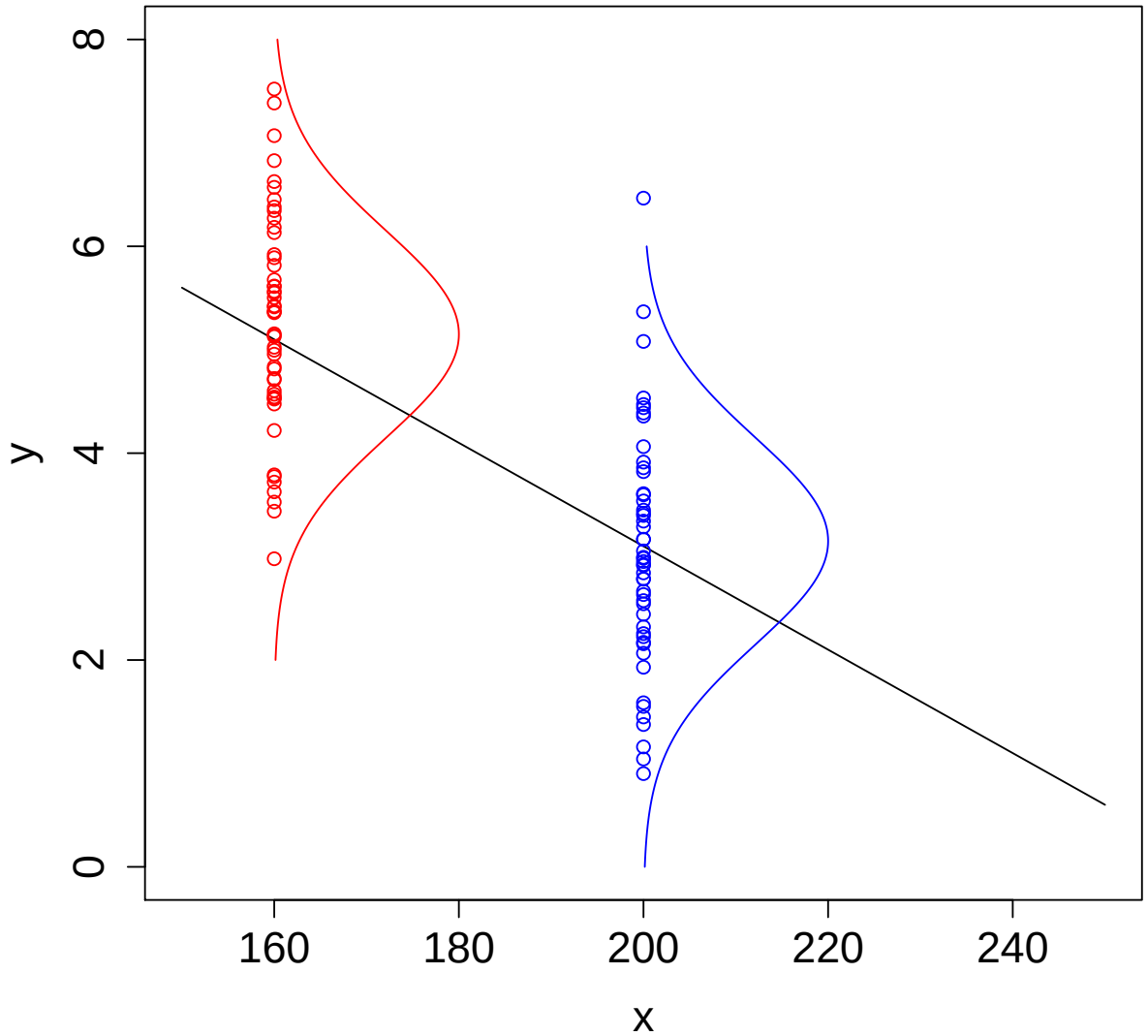
- homoscedasticiteit: bij vaste X heeft Y constante variantie

$$\mathrm{Var}(Y|X) = \mathrm{E}\left[\{Y - g(x)\}^2\right] = \sigma^2$$

die we schatten als

$$MSE = \frac{\sum_{i=1}^{n} \{y_i - \hat{g}(x_i)\}^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- normaliteit: bij vaste X is Y Normaal verdeeld


y
0
2
4
6
8
160
180
200
220
240
x

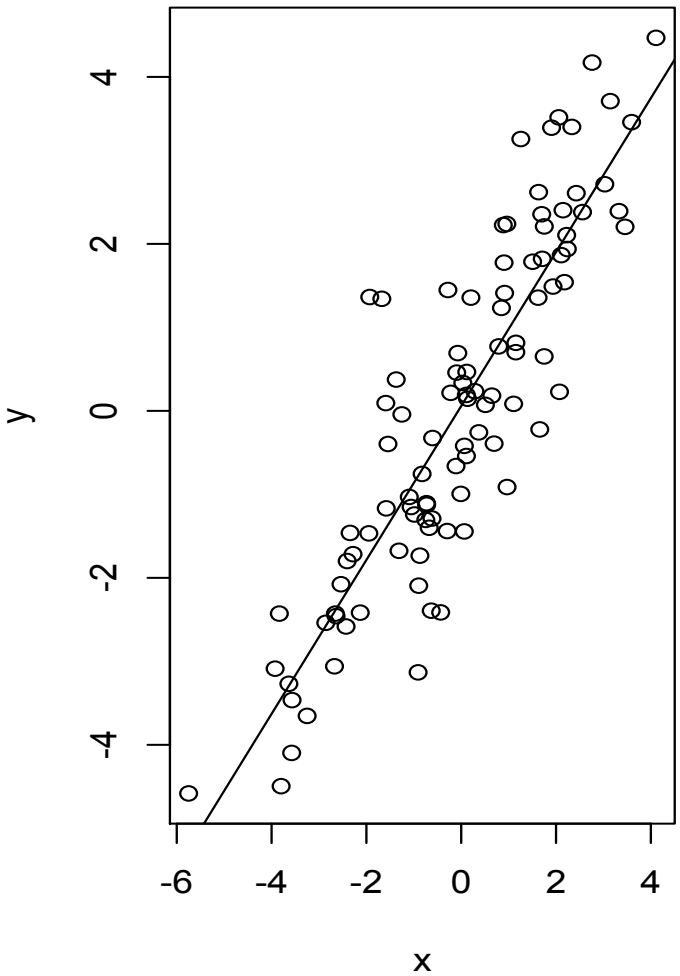
Besluitvorming voor β_1

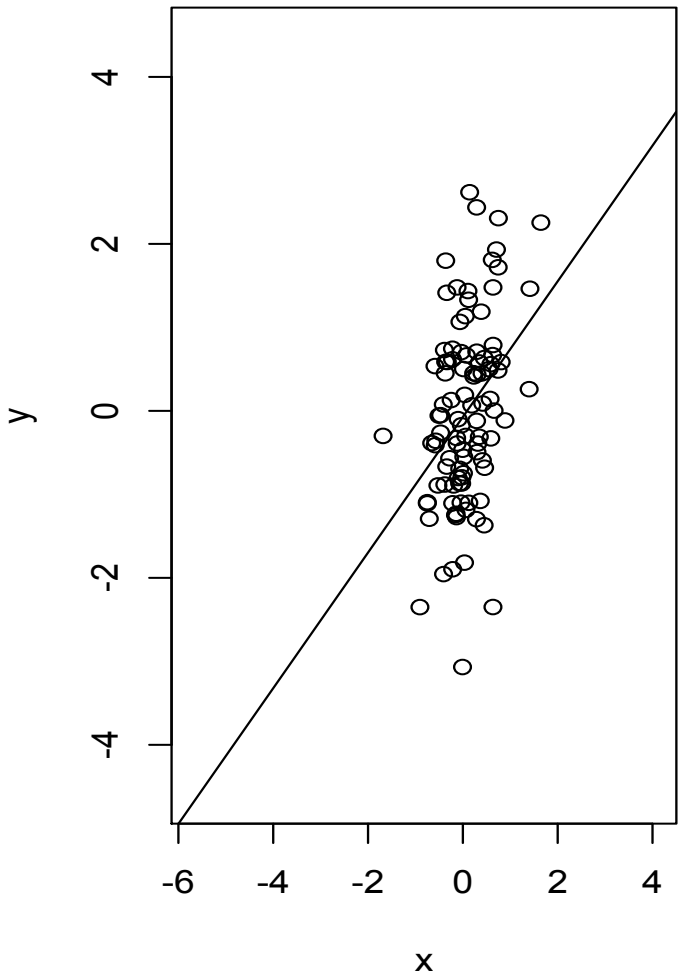
- $\hat{\beta}_1$ is onvertekende schatter van β_1.
- Standaard error van $\hat{\beta}_1$ is

$$SE(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum_i (X_i - \bar{X})^2}}$$

- Hoge spreiding op X bevordert precisie.

Spreiding en precisie





Associatie S100A8 en ESR1 expressie

Toetsen en betrouwbaarheidsintervallen (BI) voor β_1 steunen op

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

```
> summary(model)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 208.47145   28.57207   7.296 7.56e-08 ***
ESR1         -0.05926    0.01212  -4.891 4.08e-05 ***
```

p-waarde bij toets $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ is 0.0000408 (d.i. kans dat t_{29}-verdeelde veranderlijke in absolute waarde groter is dan 4.891, tweezijdig alternatief)

Associatie S100A8 en ESR1 expressie

Toetsen en betrouwbaarheidsintervallen (BI) voor β_1 steunen op

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

```
> summary(model)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 208.47145   28.57207   7.296 7.56e-08 ***
ESR1         -0.05926    0.01212  -4.891 4.08e-05 ***
```

95% BI voor β_1 vereist $t_{27,0.975} = 2.05$ (in R: `> qt(0.975,27)`)

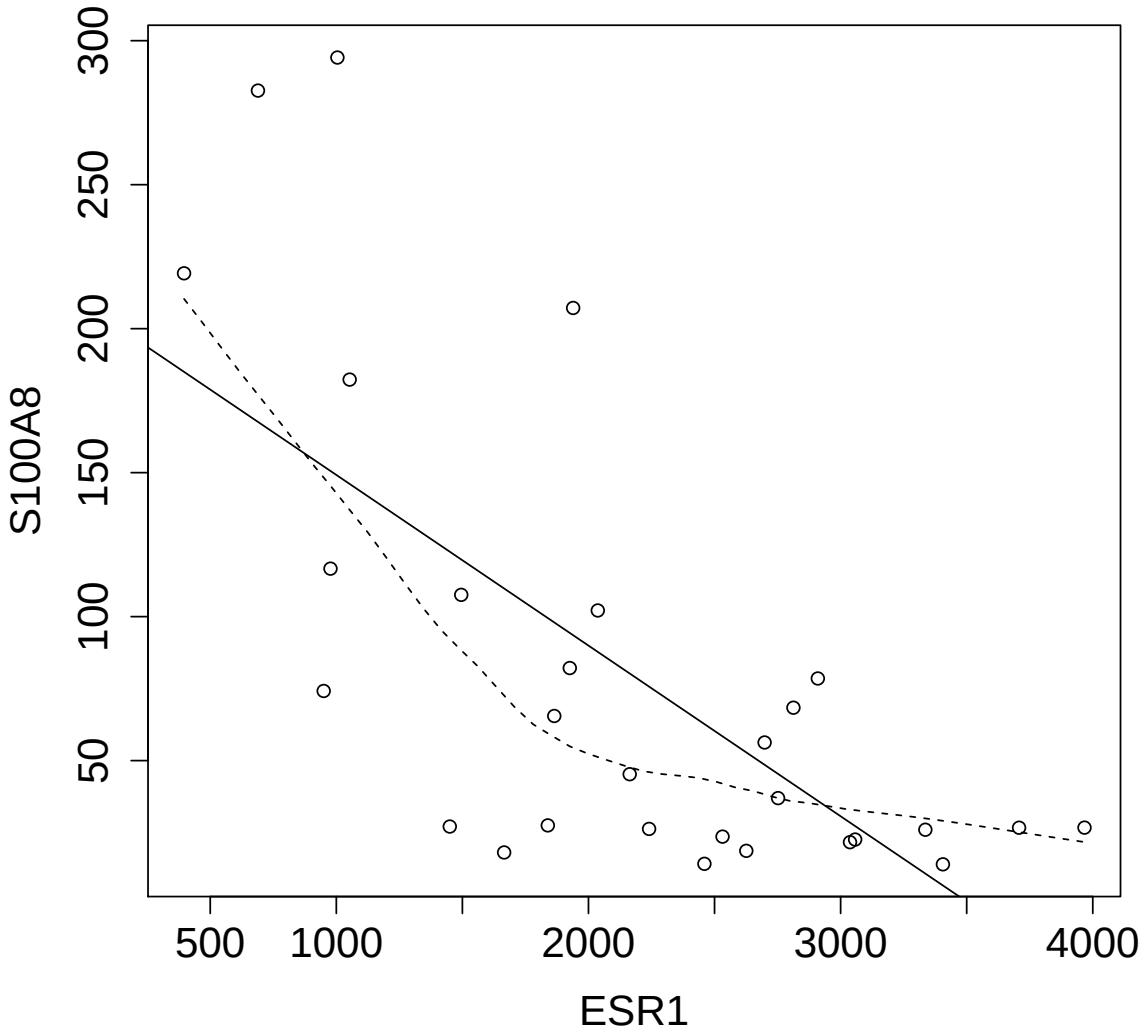
$$[-0.0593 - 2.05 \times 0.0121, -0.0593 + 2.05 \times 0.0121] = [-0.084, -0.034]$$

```
> confint(model)
                  2.5 %       97.5 %
(Intercept) 149.84639096 267.09649989
ESR1         -0.08412397  -0.03440378
```

Nagaan van veronderstellingen

- Onafhankelijkheid: design
- Lineariteit: besluitvorming geen zin als model niet lineair is
- Homoscedasticiteit: besluitvorming/p-waarde is niet betrouwbaar als de data niet homoscedastisch zijn
- Normaliteit: besluitvorming/p-waarde is niet betrouwbaar als de data niet normaal verdeeld zijn

Onderstelling van lineariteit nagaan (1)



Onderstelling van lineariteit nagaan (2)

- Een alternatief dat handiger zal blijken als er meerdere predictoren zijn, is een residuplot.

Residu's

zijn predictiefouten

$$e_i = y_i - \hat{g}(x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Onderstelling van lineariteit nagaan (2)

- Een alternatief dat handiger zal blijken als er meerdere predictoren zijn, is een residuplot.

Residu's

zijn predictiefouten

$$e_i = y_i - \hat{g}(x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Als lineaire model correct is, dan toont een scatterplot van e_i versus x_i of predicties $\hat{\beta}_0 + \hat{\beta}_1 x_i$ geen verband, anders mogelijks wel.

Onderstelling van lineariteit nagaan (2)

- Een alternatief dat handiger zal blijken als er meerdere predictoren zijn, is een residuplot.

Residu's

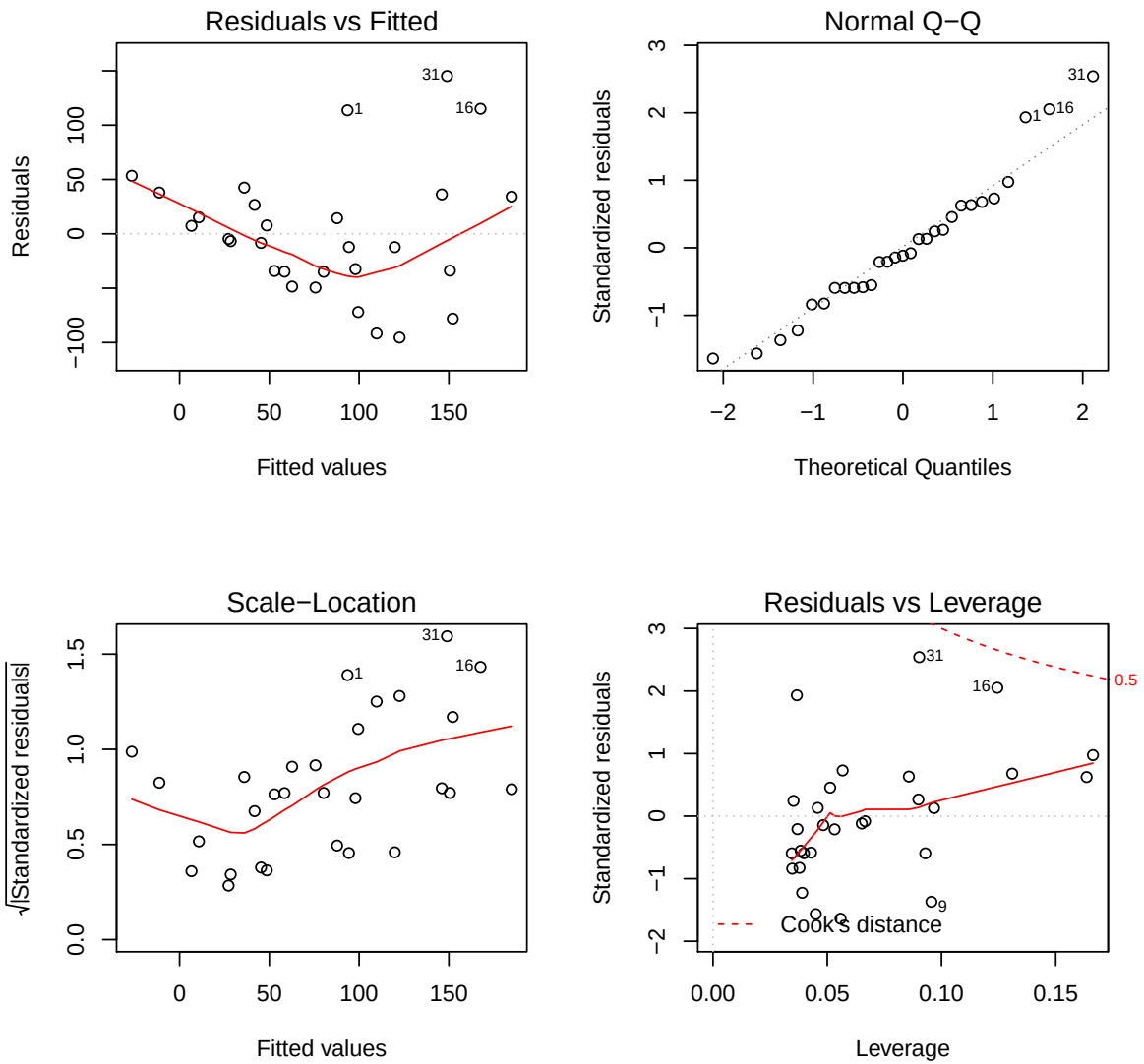
zijn predictiefouten

$$e_i = y_i - \hat{g}(x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Als lineaire model correct is, dan toont een scatterplot van e_i versus x_i of predicties $\hat{\beta}_0 + \hat{\beta}_1 x_i$ geen verband, anders mogelijks wel.

```
> par( mfrow = c(2,2))
> plot(model)
```

Lineariteit?



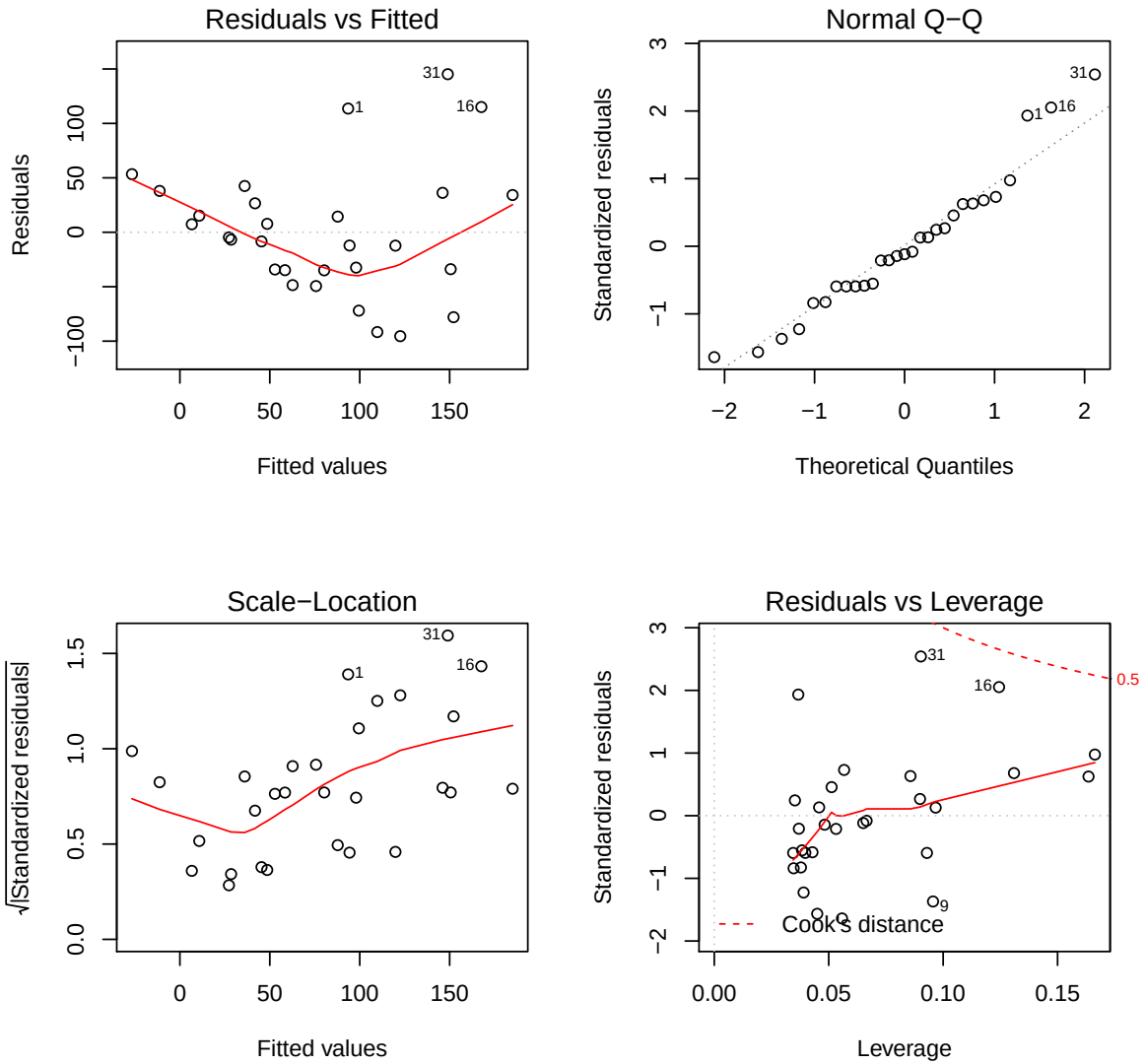
Onderstelling van homoscedasticiteit

- Residuen en kwadratische residu's dragen informatie over residuele variabiliteit.
- Als deze geassocieerd zijn met de verklarende variabelen, dan is er indicatie van heteroscedasticiteit.

$\rightarrow$ Scatterplot van of e_i versus x_i of predicties.

$\rightarrow$ Scatterplot van of e_i^2 of $\sqrt{|e_i|}$ versus x_i of predicties.

Homoscedasticiteit?



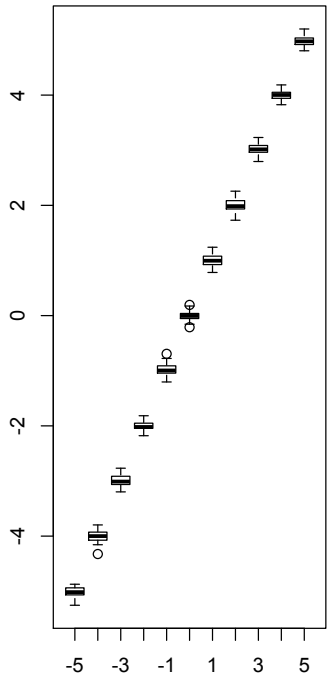
Onderstelling van Normaliteit

- Indien voldoende gegevens, zijn schatters t-verdeeld zelfs wanneer observaties niet Normaal verdeeld zijn: centrale limiet stelling
- Wat 'voldoende observaties' zijn, hangt af van hoe goed de verdeling op de Normale lijkt.

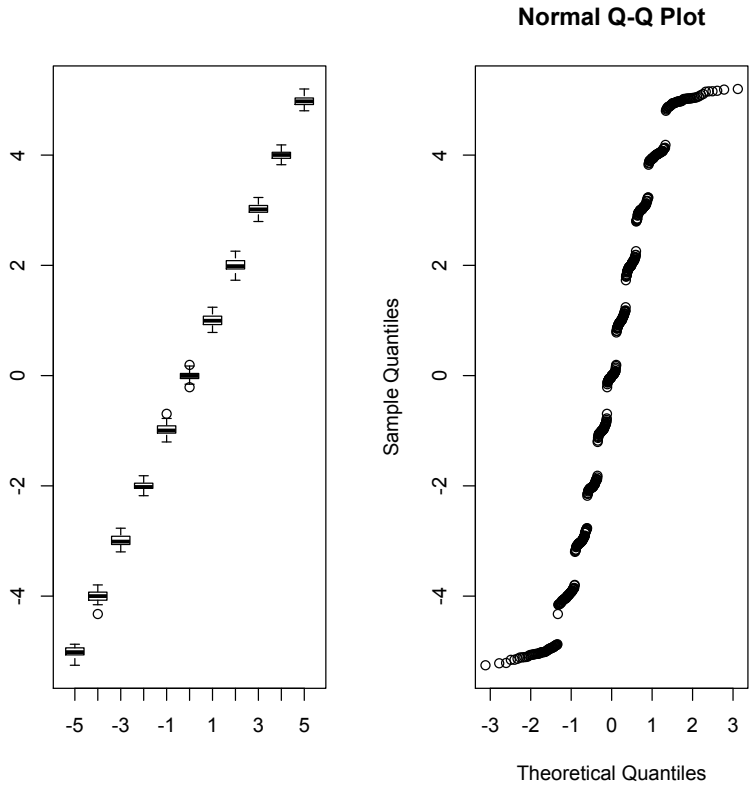
Onderstelling van Normaliteit

- Indien voldoende gegevens, zijn schatters t-verdeeld zelfs wanneer observaties niet Normaal verdeeld zijn: centrale limiet stelling
- Wat 'voldoende observaties' zijn, hangt af van hoe goed de verdeling op de Normale lijkt.
- Onderstelling is dat uitkomsten Normaal verdeeld zijn bij vaste waarden van de verklarende variabelen.
- QQ-plot van de uitkomsten is misleidend.
- QQ-plot van de residu's is zinvol: als de residu's niet Normaal verdeeld zijn, dan ook de uitkomsten niet bij gegeven X-waarden.

Normaliteit checken

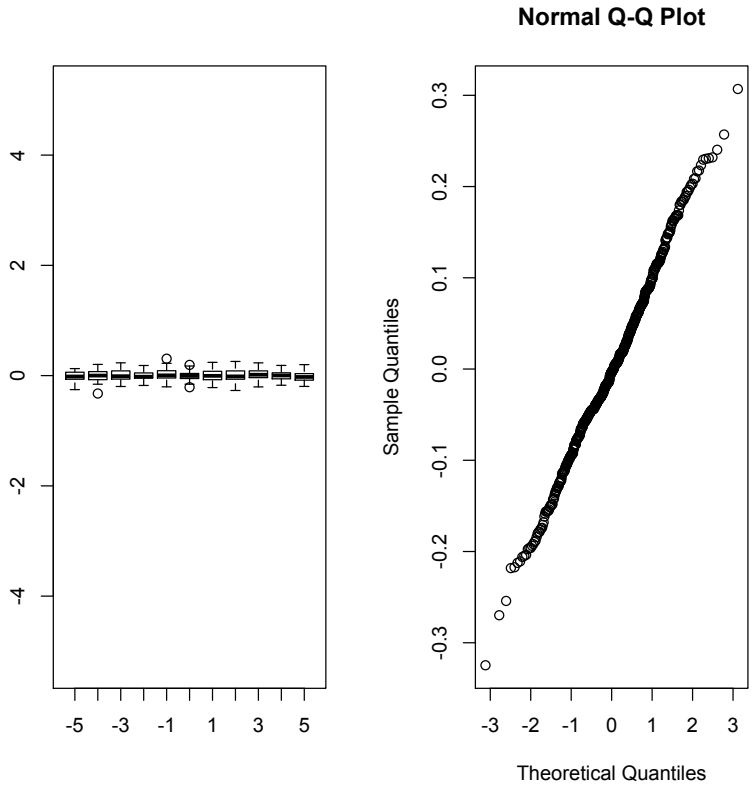


Normaliteit checken



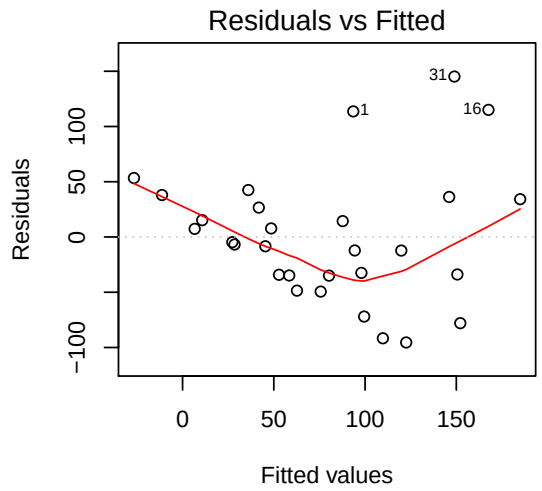
De uitkomst is niet i.i.d. normaal verdeeld: alle Y_i hebben immers een ander gemiddelde $\mu_i = E[Y_i|X_i] \rightarrow Y_i \sim N(\beta_0+\beta_1 X_i, \sigma^2)$

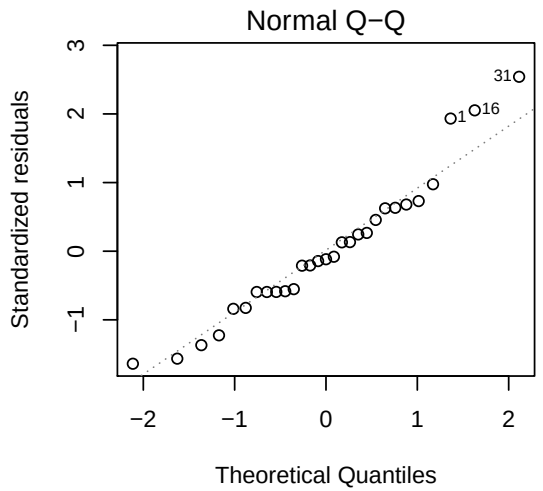
Normaliteit checken

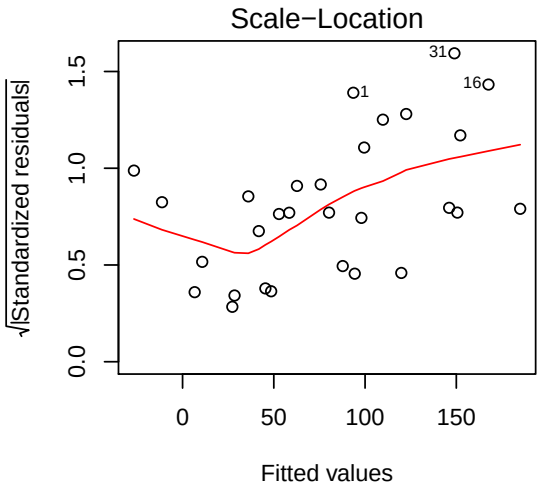


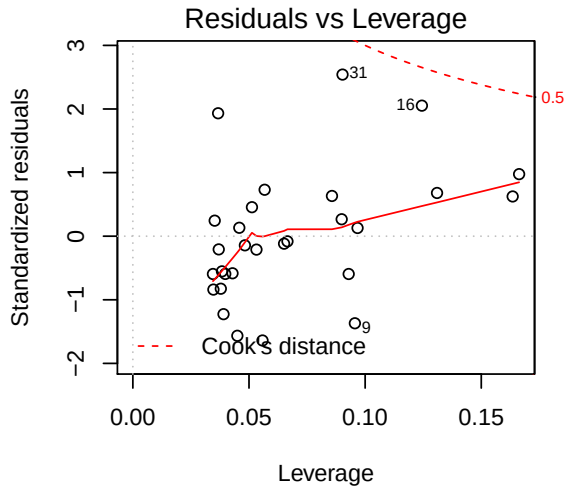
De residuelen zijn wel i.i.d. normaal verdeeld: $\epsilon_i \sim N(0, \sigma^2)$

Normaliteit?









Wat als lineariteit, homoscedasticiteit of normaliteit vals is?

- Transformatie van afhankelijke variabele kan helpen om normaliteit en homoscedasticiteit te bekomen.

voorbeeld

$\sqrt{Y}, Y^2, 1/Y, \exp Y, \exp -Y, \ln Y$

Wat als lineariteit, homoscedasticiteit of normaliteit vals is?

- Transformatie van afhankelijke variabele kan helpen om normaliteit en homoscedasticiteit te bekomen.

voorbeeld

$\sqrt{Y}, Y^2, 1/Y, \exp Y, \exp -Y, \ln Y$

- Transformatie van onafhankelijke veranderlijke wijzigt de verdeling van Y bij gegeven X niet:
  - helpt niet om normaliteit of homoscedasticiteit te bekomen;
  - helpt wel om lineariteit te bekomen wanneer er normaliteit en homoscedasticiteit is.

Wat als homoscedasticiteit of normaliteit vals is?

- Vaak wordt heteroscedasticiteit of niet-normaliteit veroorzaakt doordat uitkomst slechts waarden over beperkt gebied kan aannemen en lineaire regressiemodel daardoor niet geschikt is
- Oplossing: uitkomst zó transformeren dat ze alle reële waarden kan aannemen.

voorbeeld

ln Y voor positieve uitkomsten Y

Wat als homoscedasticiteit of normaliteit vals is?

- Vaak wordt heteroscedasticiteit of niet-normaliteit veroorzaakt doordat uitkomst slechts waarden over beperkt gebied kan aannemen en lineaire regressiemodel daardoor niet geschikt is
- Oplossing: uitkomst zó transformeren dat ze alle reële waarden kan aannemen.

voorbeeld

ln Y voor positieve uitkomsten Y

- Variantie-stabiliserende transformaties.

voorbeeld

arcsin $\sqrt{Y}$ voor proporties Y

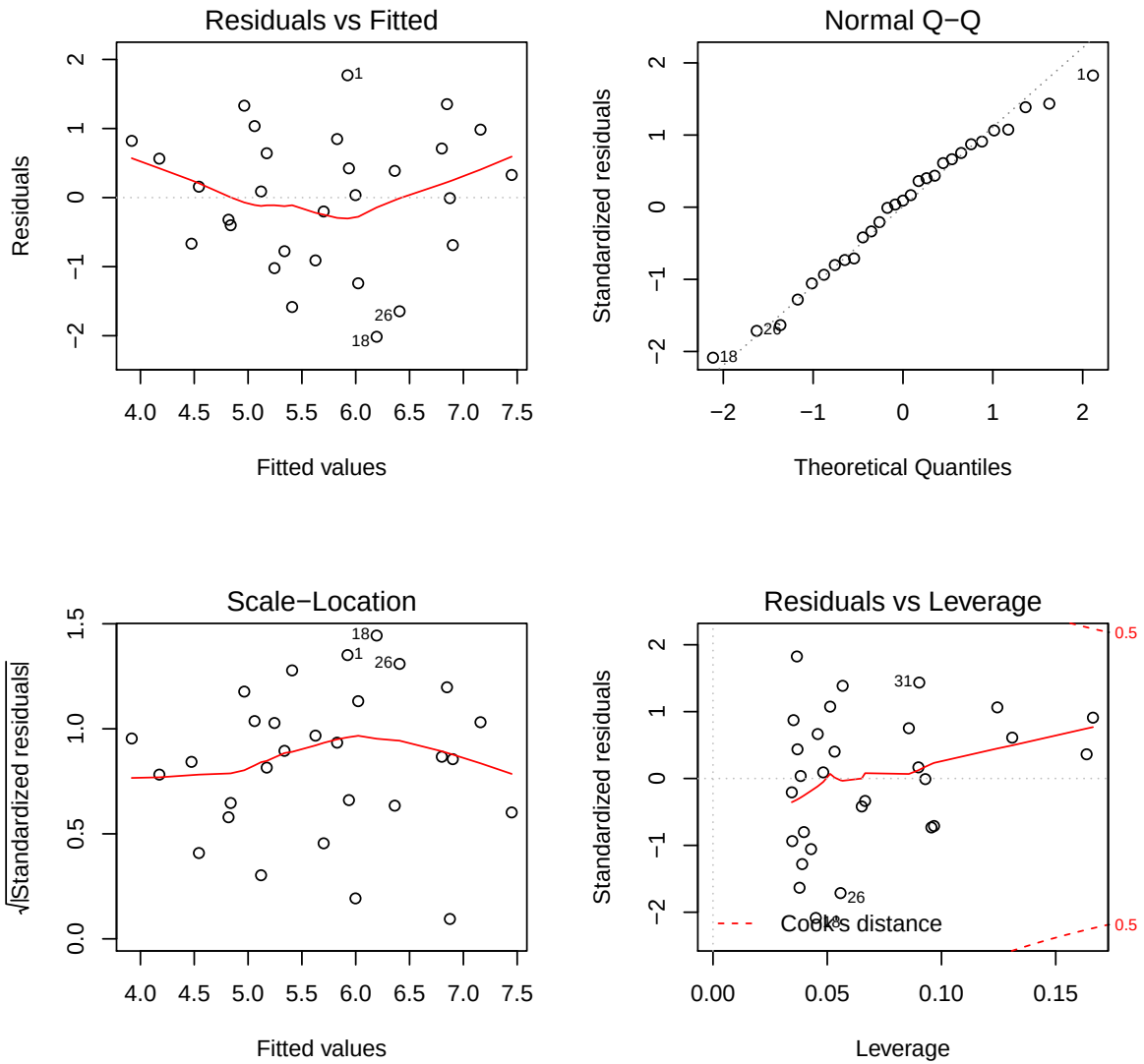
Transformatie van de uitkomst

- Heteroscedasticiteit mogelijk doordat uitkomst positief is, maar lineaire model dit niet respecteert.
- ln-transformatie maakt de uitkomst reëelwaardig.
- In genexpressie studies maakt men meestal gebruik van $\log_2$ transformatie

```
> model2 <- lm(I(log2(S100A8))
ESR1,data=subset(borstkanker,S100A8<2000)))
> summary(model2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.8402230  0.4715434  16.627 1.04e-15 ***
ESR1        -0.0009883  0.0002000  -4.943 3.55e-05 ***
```

Residuplots



Wat als onderstelling van lineariteit vals is?

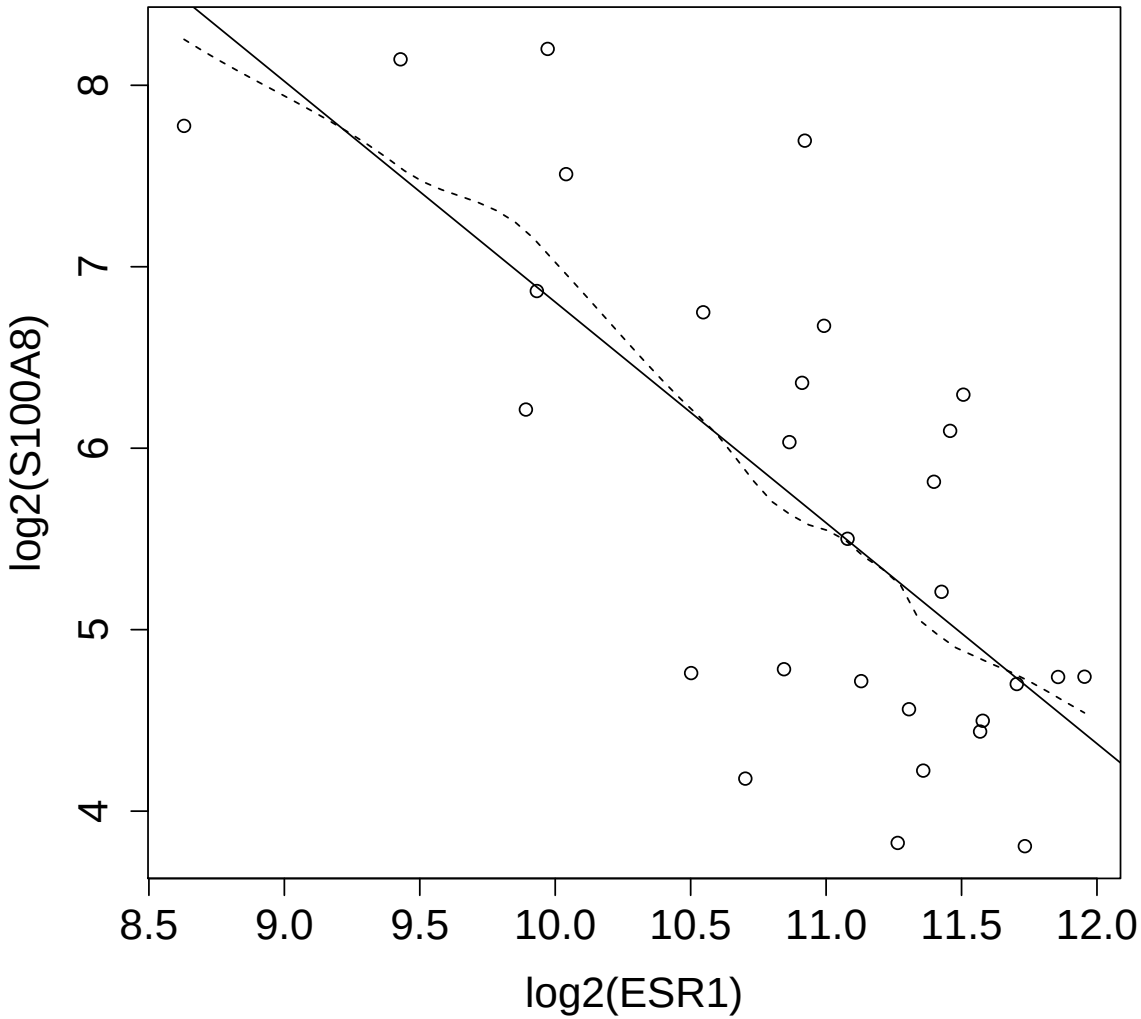
- Transformatie van de afhankelijke variabele.
- Transformatie van de onafhankelijke variabele.

voorbeeld

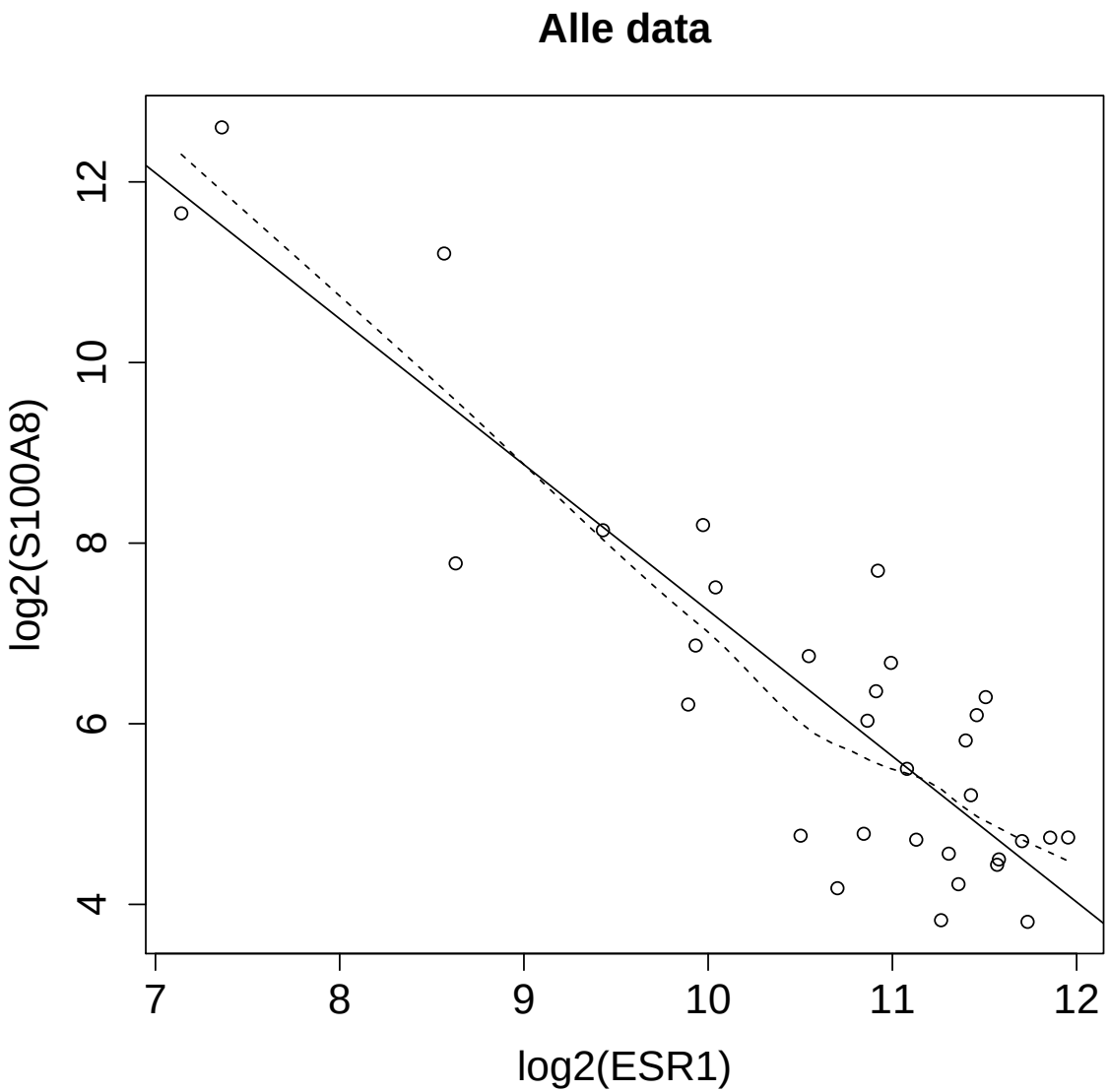
Als afhankelijke of onafhankelijke variabele scheef verdeeld is naar rechts, helpt log-transformatie vaak.

- Op originele plot zagen we inderdaad een soort exponentiële trend. Lijkt logisch om beide genexpressies log te transformeren
- Intensiteitsmetingen zijn vaak log-normaal verdeeld.

Associatie tussen S100A8 en ESR1 expressie op $\log_2$ schaal



Associatie tussen S100A8 en ESR1 expressie op $\log_2$ schaal

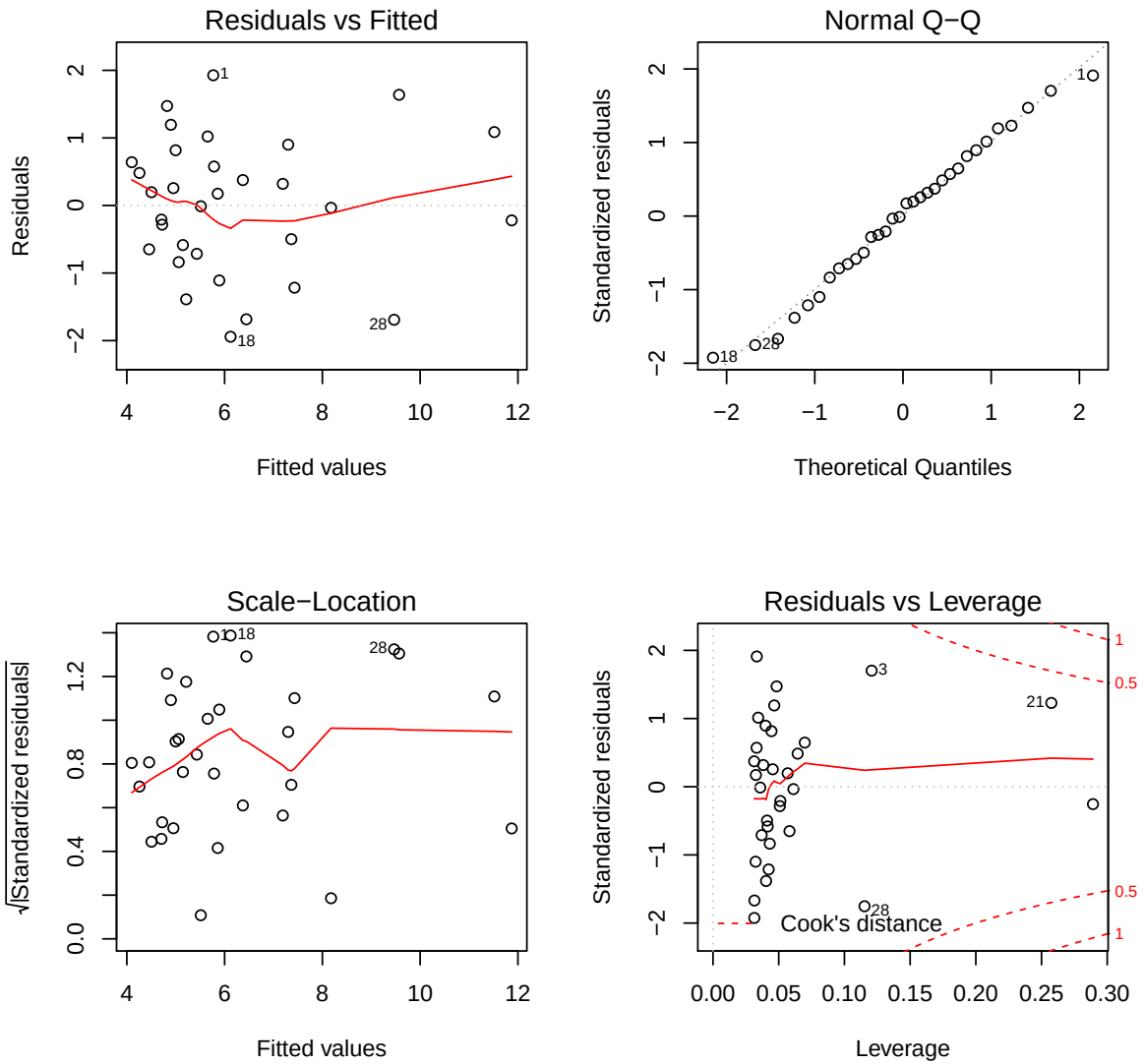


log-log regressie

Transformatie van predictor en response variable in formule geeft problemen als je de plot functie aanroep op model $\rightarrow$ daarom log-transformeren we expliciet.

```
> borstkanker$logS100A8 <- log2(borstkanker$S100A8)
> borstkanker$logESR1 <- log2(borstkanker$ESR1)
> model3=lm(logS100A8~logESR1,borstkanker)
> summary(model3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.401      1.603   14.60 3.57e-15 ***
logESR1       -1.615      0.150  -10.76 8.07e-12 ***
---
```

Residuplots



Interpretatie van parameters in een log-log model

- Een groep patiënten met een ESR1 expressie die 1 eenheid op de $\log_2$ schaal hoger ligt dan dat van een andere groep patiënten heeft gemiddeld gezien een expressie-niveau van S100A8 dat 1.6 eenheden lager ligt (95% BI [-1.9 -1.3]).

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \log \text{ESR}_1, \quad \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \log \text{ESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1) = -1.615$$

Interpretatie van parameters in een log-log model

- Een groep patiënten met een ESR1 expressie die 1 eenheid op de $\log_2$ schaal hoger ligt dan dat van een andere groep patiënten heeft gemiddeld gezien een expressie-niveau van S100A8 dat 1.6 eenheden lager ligt (95% BI [-1.9 -1.3]).

```
> confint(model3)
                  2.5 %    97.5 %
(Intercept) 20.128645 26.674023
logESR1     -1.921047 -1.308185
```

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \quad \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1) = -1.615$$

Interpretatie van parameters in een log-log model

- Een groep patiënten met een dubbel zo hoge ESR1 expressie hebben gemiddeld een S100A8 expressie die 3.1 keer lager ligt (95% BI [2.5,3.8]).

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \log \text{ESR}_1, \quad \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \log \text{ESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1)$$

$$\log_2 \left[\frac{\hat{\mu}_2}{\hat{\mu}_1}\right] = -1.615 \log_2 \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]$$

$$\frac{\hat{\mu}_2}{\hat{\mu}_1} = \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]^{-1.615} = 2^{-1.615} = 0.3264649$$

Interpretatie van parameters in een log-log model

- Een groep patiënten met een dubbel zo hoge ESR1 expressie hebben gemiddeld een S100A8 expressie die 3.1 keer lager ligt (95% BI [2.5,3.8]).

```
> 1/(2^confint(model3))
                      2.5 %       97.5 %
(Intercept) 8.723163e-07 9.339397e-09
logESR1     3.786977e+00 2.476298e+00
> 1/(2^coef(model3)[2])
logESR1
 3.0623
```

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \quad \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1)$$

$$\log_2 \left[\frac{\hat{\mu}_2}{\hat{\mu}_1}\right] = -1.615 \log_2 \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]$$

$$\frac{\hat{\mu}_2}{\hat{\mu}_1} = \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]^{-1.615} = 2^{-1.615} = 0.3264649$$

Interpretatie van parameters in een log-log model

- Een groep patiënten met een ESR1 expressie die 1% hoger ligt dan dat van een andere groep patiënten heeft gemiddeld gezien een expressie-niveau van S100A8 dat ongeveer 1.6% lager ligt (95% BI [-1.9% -1.3%]).

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \quad \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1)$$

$$\log_2 \left[\frac{\hat{\mu}_2}{\hat{\mu}_1}\right] = -1.615 \log_2 \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]$$

$$\frac{\hat{\mu}_2}{\hat{\mu}_1} = \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]^{-1.615} = 1.01^{-1.615} = 0.984 \approx -1.6\%$$

Interpretatie van parameters in een log-log model

- Een groep patiënten met een ESR1 expressie die 1% hoger ligt dan dat van een andere groep patiënten heeft gemiddeld gezien een expressie-niveau van S100A8 dat ongeveer 1.6% lager ligt (95% BI [-1.9% -1.3%]).

```
> confint(model3)
                  2.5 %    97.5 %
(Intercept) 20.128645 26.674023
logESR1     -1.921047 -1.308185
```

$$\log_2 \hat{\mu}_1 = 23.401 - 1.615 \times \text{logESR}_1, \quad \log_2 \hat{\mu}_2 = 23.401 - 1.615 \times \text{logESR}_2$$

$$\log_2 \hat{\mu}_2 - \log_2 \hat{\mu}_1 = -1.615(\log_2 \text{ESR}_2 - \log_2 \text{ESR}_1)$$

$$\log_2 \left[\frac{\hat{\mu}_2}{\hat{\mu}_1}\right] = -1.615 \log_2 \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]$$

$$\frac{\hat{\mu}_2}{\hat{\mu}_1} = \left[\frac{\text{ESR}_2}{\text{ESR}_1}\right]^{-1.615} = 1.01^{-1.615} = 0.984 \approx -1.6\%$$

Besluitvorming voor gemiddelde uitkomst

- $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is onvertekende schatter van $E(Y|X = x) = \beta_0 + \beta_1 x$.
- Standaard error van $\hat{g}(x)$ is

$$SE_{\hat{g}(x)} = \sqrt{MSE \left\{ \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right\}}$$

- Predicties meest precies in $x = \bar{X}$ en zelfs even precies dan wanneer alle $X = X_1 \ldots X_n$ in de steekproef gelijk zouden zijn.

Besluitvorming voor gemiddelde uitkomst

- $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is onvertekende schatter van $E(Y|X = x) = \beta_0 + \beta_1 x$.
- Standaard error van $\hat{g}(x)$ is

$$SE_{\hat{g}(x)} = \sqrt{MSE \left\{ \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right\}}$$

- Predicties meest precies in $x = \bar{X}$ en zelfs even precies dan wanneer alle $X = X_1 \ldots X_n$ in de steekproef gelijk zouden zijn.
- Toetsen en BI voor $E(Y|X = x)$ steunen op statistiek

$$\frac{\hat{g}(x) - g(x)}{SE_{\hat{g}(x)}} \sim t_{n-2}$$

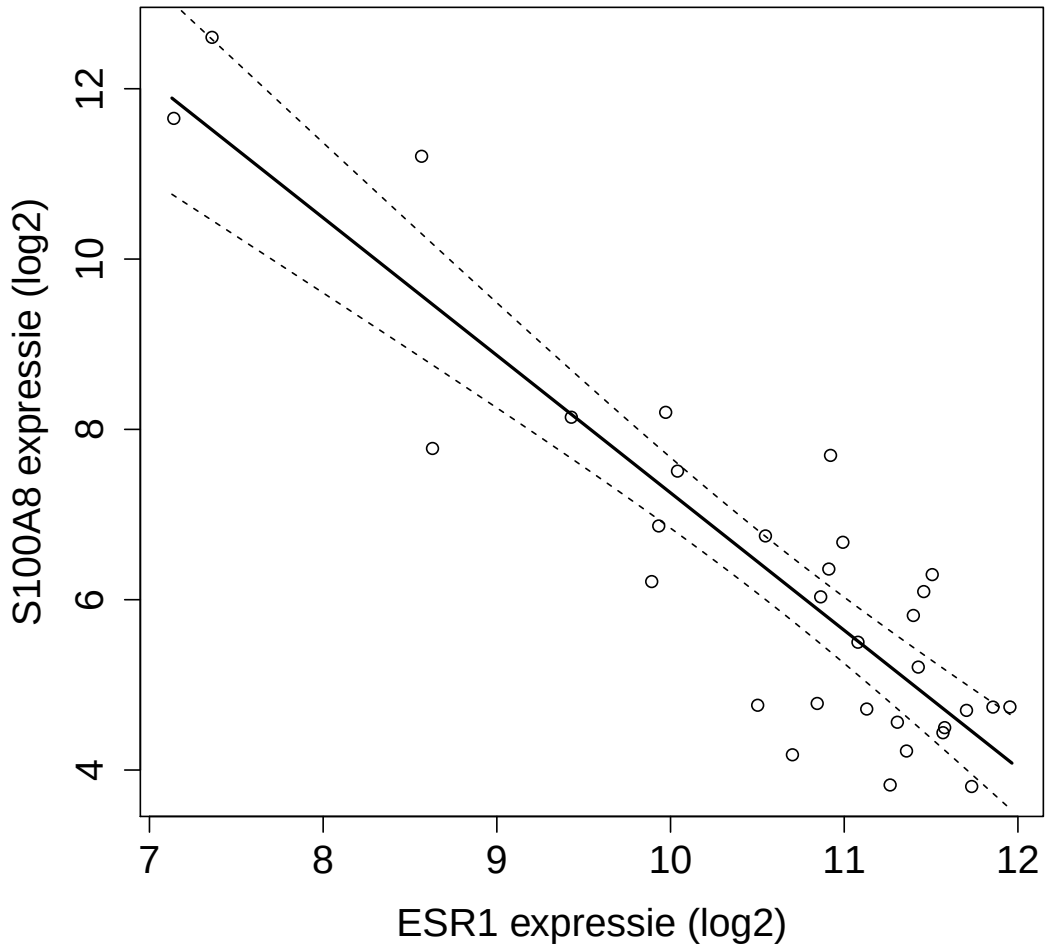
die t-verdeling volgt met n-2 vrijheidsgraden.

Verwachte uitkomst in R

```
> g<-predict(model3,newdata=data.frame(logESR1=log2(140:4000)),
+ interval="confidence")
> g
            fit       lwr       upr
1     11.890282 10.760819 13.019744
          .         .         .
          .         .         .
3861   4.081190  3.525000  4.637380
```

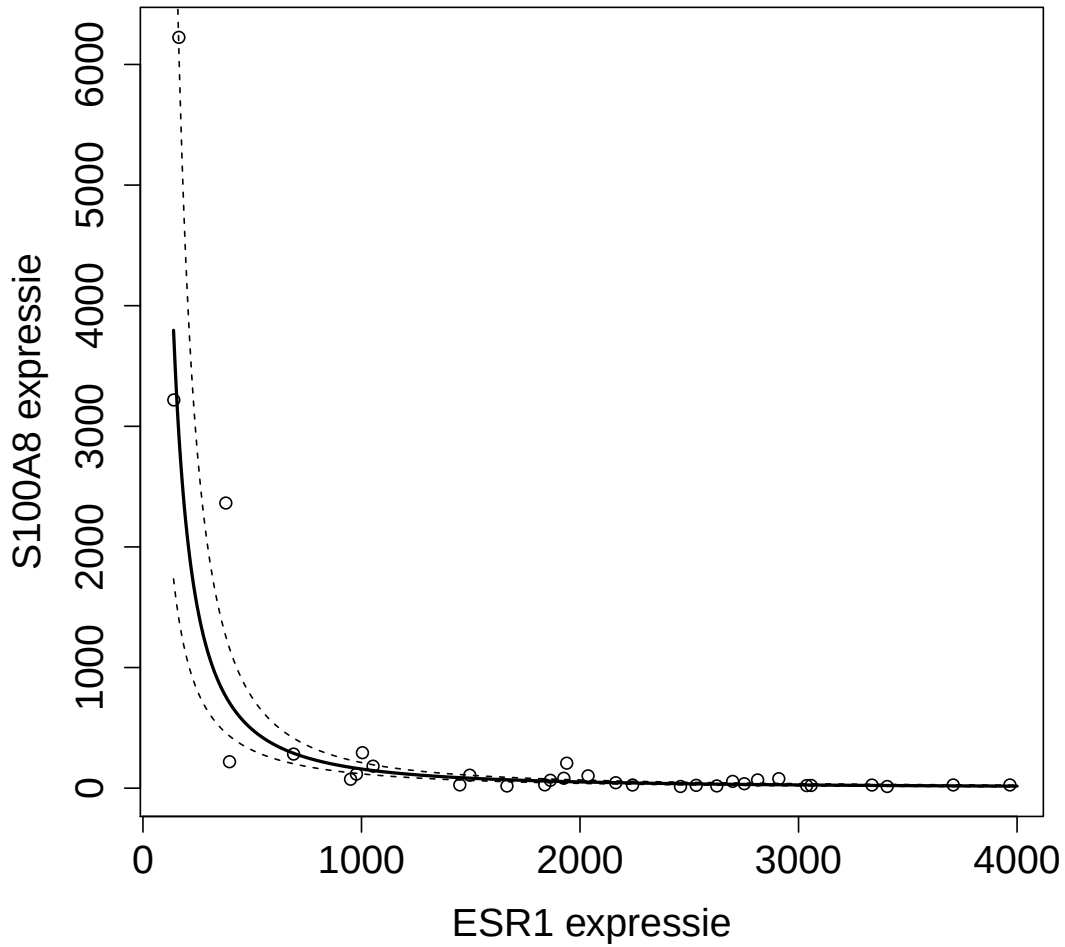
Verwachte uitkomst met 95% BI

```
> plot(logS100A8~logESR1,borstkanker,xlab="ESR1 expressie (log2)",ylab="S100A8 expressie (log2)")
> matplot(log2(140:4000),g,lty=c(1,2,2),lwd=c(2,1,1),type="l",add=TRUE,col=1)
```



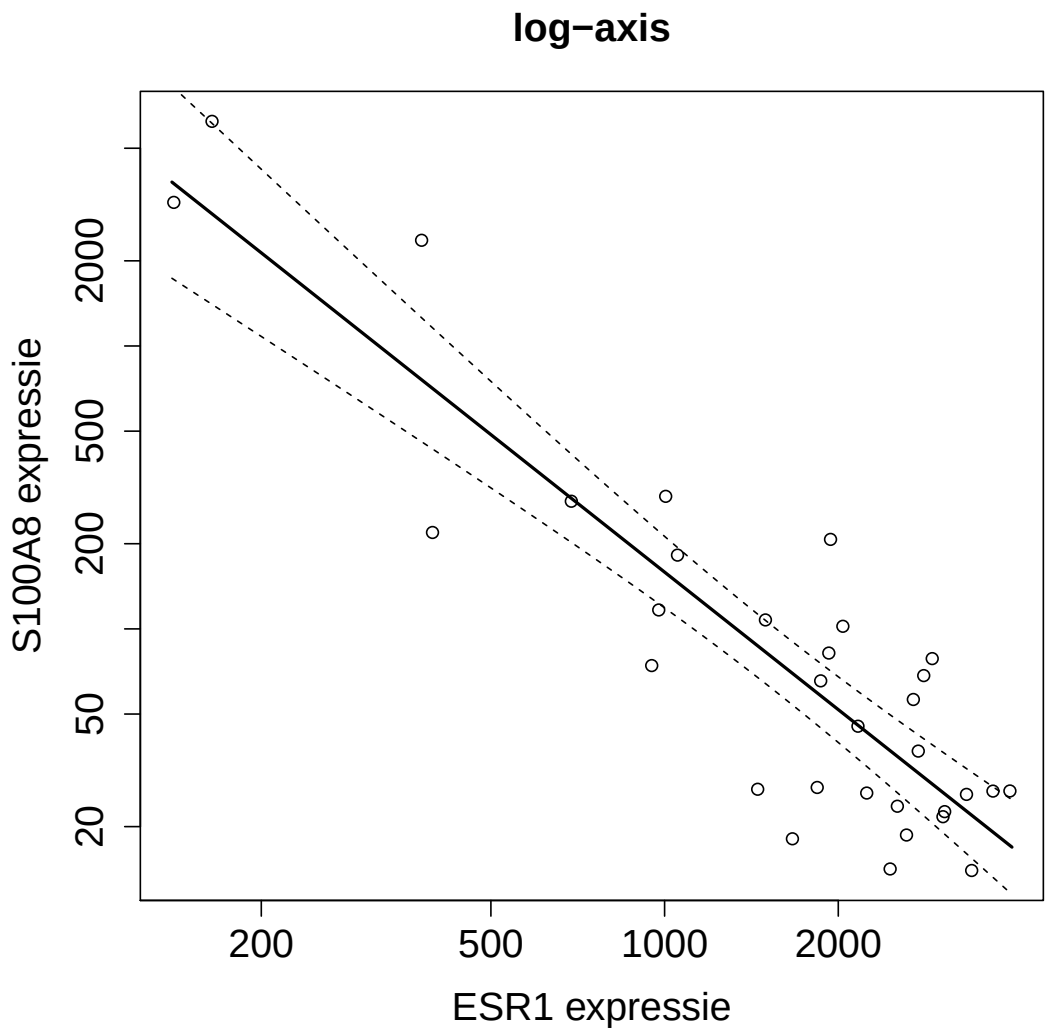
Verwachte uitkomst met 95% BI

```
> plot(S100A8~ESR1,borstkanker,xlab="ESR1 expressie",ylab="S100A8 expressie")
> matplot(140:4000,2^g,lty=c(1,2,2),lwd=c(2,1,1),type="l",add=TRUE,col=1)
```



Verwachte uitkomst met 95% BI

```
> plot(S100A8~ESR1,borstkanker,xlab="ESR1 expressie",ylab="S100A8 expressie",log="xy",main="log-axis")
> matplot(140:4000,2^g,lty=c(1,2,2),lwd=c(2,1,1),type="l",add=TRUE,col=1)
```



Predictie-intervallen

- Geschatte regressiemodel kan ook worden gebruikt om een predictie te maken voor één uitkomst van één experiment waarbij een nieuwe uitkomst Y^* bij een gegeven x zal geobserveerd worden.
- Het is belangrijk in te zien dat dit experiment nog moet worden uitgevoerd; we wensen dus een nog niet-geobserveerde individuele uitkomst te voorspellen.
- Aangezien Y^* een nieuwe, onafhankelijke observatie voorstelt, weten we dat

$$Y^* = g(x) + \epsilon^*$$

met $\epsilon^* \sim N(0, \sigma^2)$ en ϵ^* onafhankelijk van de steekproefobservaties $Y_1, \ldots, Y_n$.

- $\hat{g}(x)$ is een schatting van gemiddelde log-S100A8 expressie bij log-ESR1 expressie x
  dus een schatting van conditioneel gemiddelde $\mathrm{E}[Y|x]$.
- $\hat{g}(x)$ is ook goede predictie van nieuwe log-S100A8 expressiewaarde Y^* bij gegeven log-ESR1 expressieniveau x.
  - $\hat{g}(x)$ is schatting van $\mathrm{E}[Y|x]$, het punt op de regressierechte bij x.
  - Bij gegeven x worden individuele uitkomsten Y normaal verdeeld verondersteld rond $\mathrm{E}[Y|x]$
  - Normale distributie is symmetrisch dus even waarschijnlijk om uitkomst groter of kleiner dan $\mathrm{E}[Y|x]$ te observeren
  - Geen informatie over mogelijke afwijking

  $\rightarrow$ Punt op (geschatte) regressierechte is beste predictie van een individuele uitkomst bij een gegeven x.

- We voorspellen dus een nieuwe log-S100A8 meting bij een gekend log2-ESR1 expressieniveau x door

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 \times x$$

- Merk op: $\hat{y}(x)$ eigenlijk numeriek gelijk aan $\hat{g}(x)$
- Maar verschil in interpretatie: predictie vs schatting van conditioneel gemiddelde

$\rightarrow$ gebruik van andere notatie

$\rightarrow$ steekproef distributies zijn verschillend:

- SE voor geschatte gemiddelde uitkomst $\hat{g}(x)$ gedreven door de onzekerheid op $\hat{\beta}_0$ en $\hat{\beta}_1$.
- SE voor predictie $\hat{y}(x)$ gedreven door onzekerheid op het geschatte gemiddelde + bijkomende onzekerheid t.g.v. random variatie van observaties rond de het conditionele gemiddelde (de regressie rechte).

- Nieuwe observatie is onafhankelijk van observaties in steekproef dus

$$\text{SE}_{\hat{Y}(x)} = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2_{\hat{g}(x)}} = \sqrt{MSE \left\{ 1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right\}}.$$

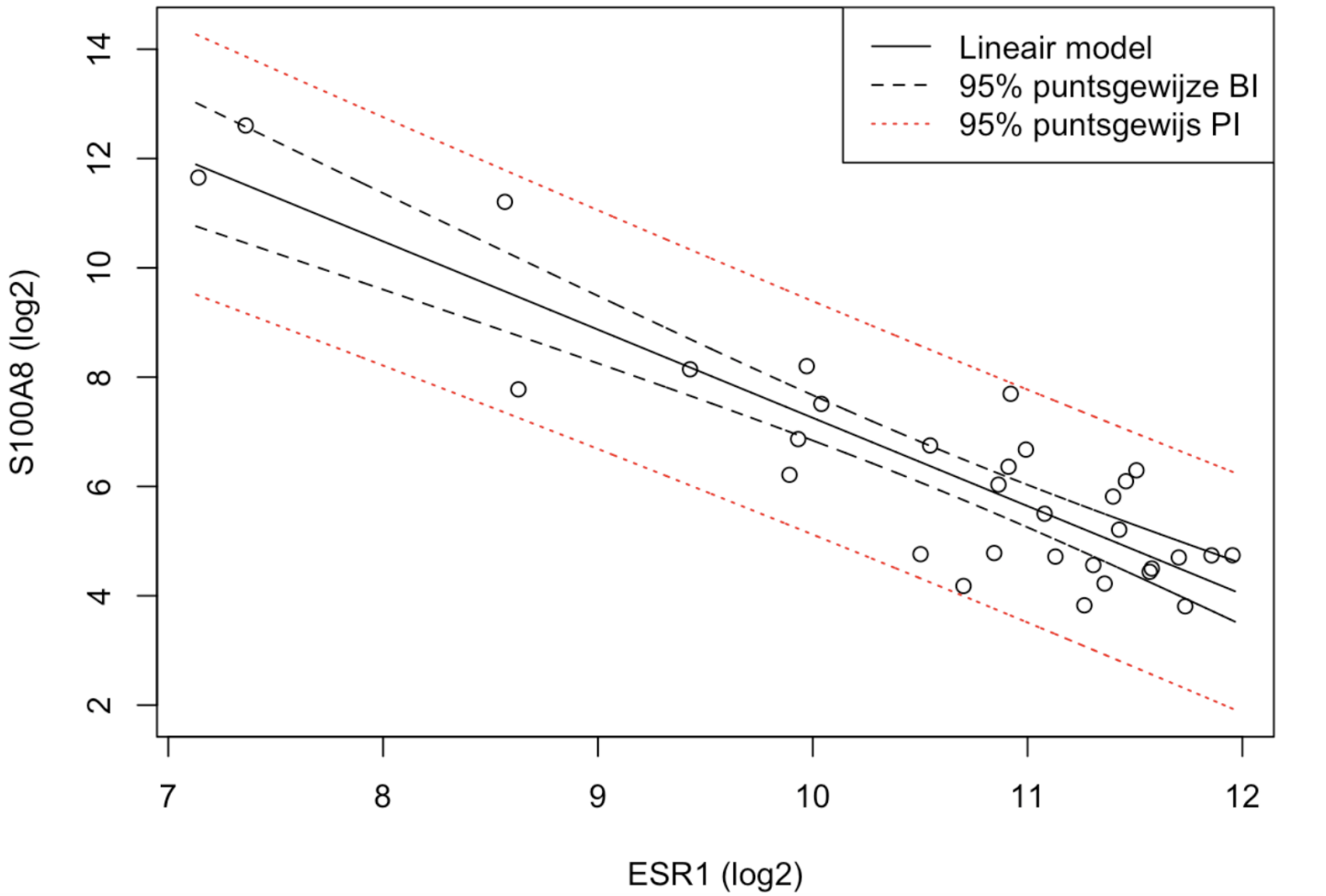
- Opnieuw

$$\frac{\hat{Y}(x) - Y}{\text{SE}_{\hat{Y}(x)}} \sim t_{n-2}$$

- Gebruik statistiek om betrouwbaarheidsinterval op de predictie te construeren: predictie-interval

Predictie-interval in R

```
> grid=log2(140:4000)
> p <- predict(lm2,newdata=data.frame(log2ESR1=grid),
+ interval="prediction")
> head(p)
       fit      lwr      upr
1 11.89028 9.510524 14.27004
2 11.87370 9.495354 14.25205
3 11.85724 9.480288 14.23419
4 11.84089 9.465324 14.21646
5 11.82466 9.450461 14.19886
6 11.80854 9.435698 14.18138
```


Lineair model
95% puntsgewijze BI
95% puntsgewijs PI
S100A8 (log2)
ESR1 (log2)
2
4
6
8
10
12
14
7
8
9
10
11
12

- Predictie-interval is een verbeterde versie van een referentie-interval wanneer de modelparameters niet gekend zijn. Houdt immers rekening met
  - Onzekerheid op het geschatte gemiddelde: gebruik van standaard error op predictie i.p.v. standaard deviatie
  - Onzekerheid op geschatte standaard deviatie (gebruik van t-verdeling i.p.v normaal verdeling).
- NHANES Voorbeeld
  - Referentie interval normale bloeddruk NHANES: [98, 142.7] mmHg
  - PI normale bloeddruk: [97.8 142.9] mmHg
  - PI niet zoveel breder: 180 observaties!

```
> lmBpNorm <- lm(bpSys~1,data=nhanesSubHealthy)
> predInt <- predict(lmBpNorm,interval="prediction",
+ newdata=data.frame(geenpredictor=1))
> round(predInt,1)
    fit  lwr   upr
1 120.4 97.8 142.9
```

Kwadratensommen en Anova

Kwadratensommen en de Anova Tabel

In deze sectie bespreken we de constructie van **kwadratensommen** die typisch in een tabel worden gegeven en die behoren tot de klassieke presentatiewijze van een regressie-analyse. De tabel wordt de **variantie-analyse tabel** of **anova tabel** genoemd.

Ontbinding van de Totale Kwadratensom

Totale kwadratensom

De totale kwadratensom is gelijk aan

$$\text{SSTot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

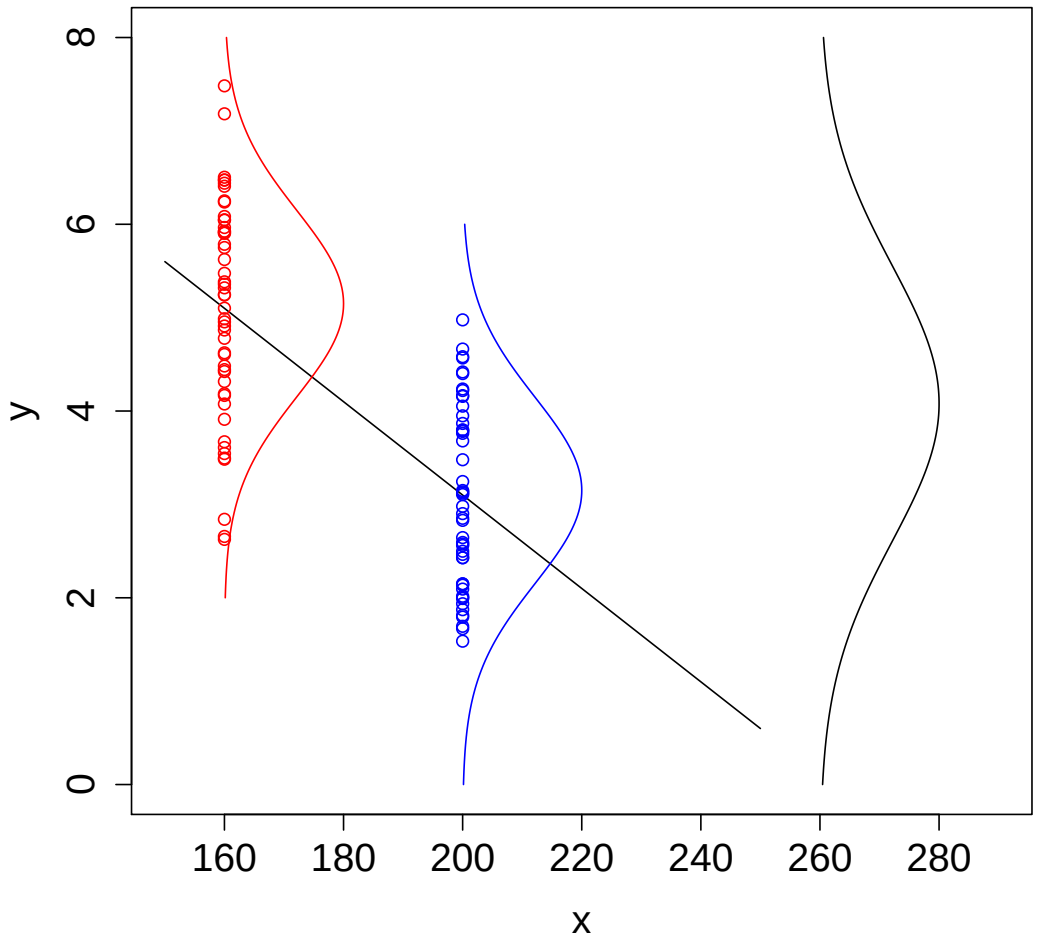
SSTot meet dus de totale variabiliteit in de uitkomstvariabele.

De statistiek

$$\frac{\text{SSTot}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

is de steekproefvariantie van de **marginale distributie** van de uitkomsten.

- In dit hoofdstuk wordt de focus hoofdzakelijk gelegd op de **conditionele distributie** van Y gegeven x
- We weten reeds dat MSE een schatter is van de variantie van de distributie van Y gegeven x.
- De **marginale distributie** van Y heeft als gemiddelde $\mathrm{E}\{Y\}$ wat geschat wordt door het steekproefgemiddelde $\bar{Y}$.
- De statistiek $\frac{\text{SSTot}}{n-1}$ is de schatter van de variantie van de marginale distributie van Y, i.e. $\mathrm{Var}[Y]$.


y
0
2
4
6
8
160
180
200
220
240
260
280
x

Kwadratensom van de regressie

De kwadratensom van de regressie is gelijk aan

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{g}(x_i) - \bar{Y})^2.$$

SSR is een maat voor de afwijking tussen de geschatte regressierechte en het steekproefgemiddelde van de uitkomsten.

Het kan ook geïnterpreteerd worden als een maat voor de afwijking tussen de geschatte regressierechte $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ en een "geschatte regressierechte" waarbij de regressor geen effect heeft op de gemiddelde uitkomst.

Deze laatste is dus eigenlijk een schatting van de regressierechte $g(x) = \beta_0$, waarin β_0 geschat wordt door $\bar{Y}$.

Anders geformuleerd: SSR meet de grootte van het regressie-effect zodat SSR ≈ 0 duidt op geen effect van de regressor en SSR > 0 duidt op een effect van de regressor.

Tenslotte herhalen we de kwadratensom van de fout:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \{Y_i - \hat{g}(x_i)\}^2 .$$

Van SSE weten we reeds dat het een maat is voor de afwijking tussen de observaties en de predicties bij de geobserveerde x_i uit de steekproef. Hoe kleiner SSE, hoe beter de fit (schatting) van de regressierechte voor predictiedoeleinden.

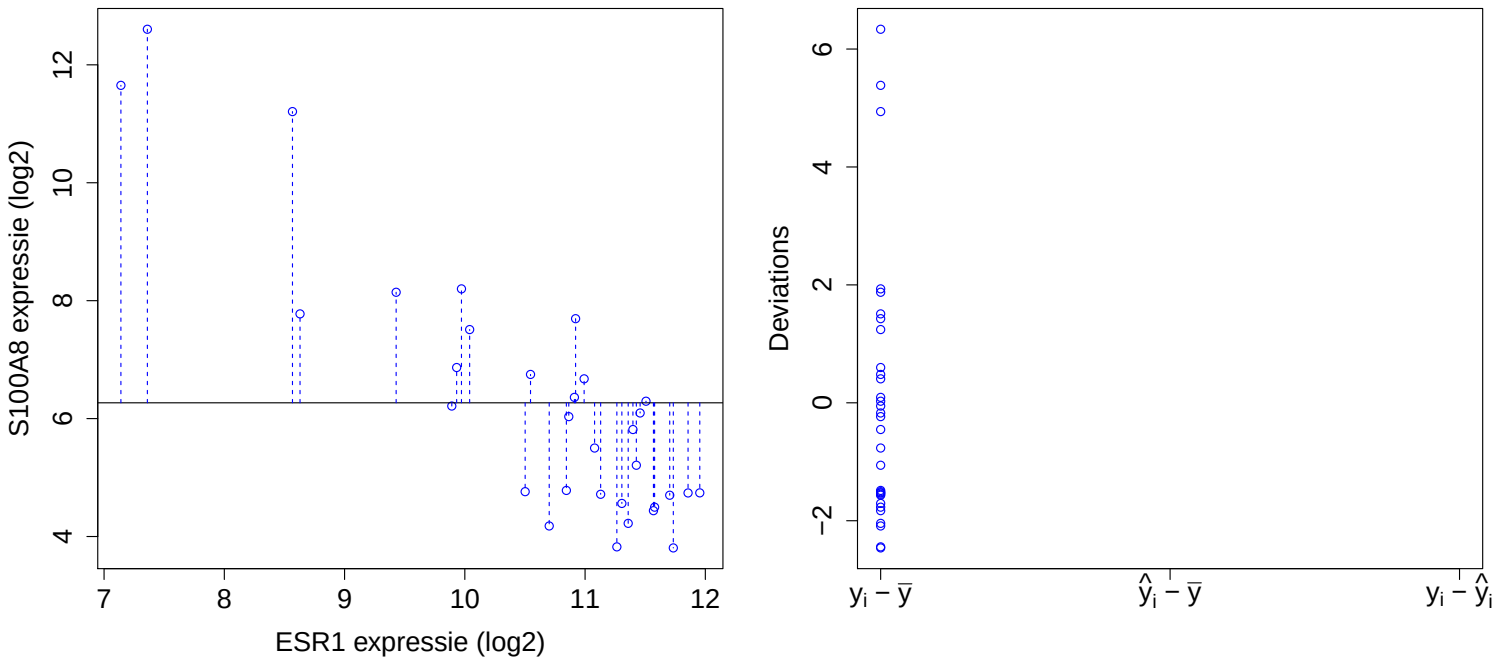
Ontbinding totale kwadratensom

Er geldt

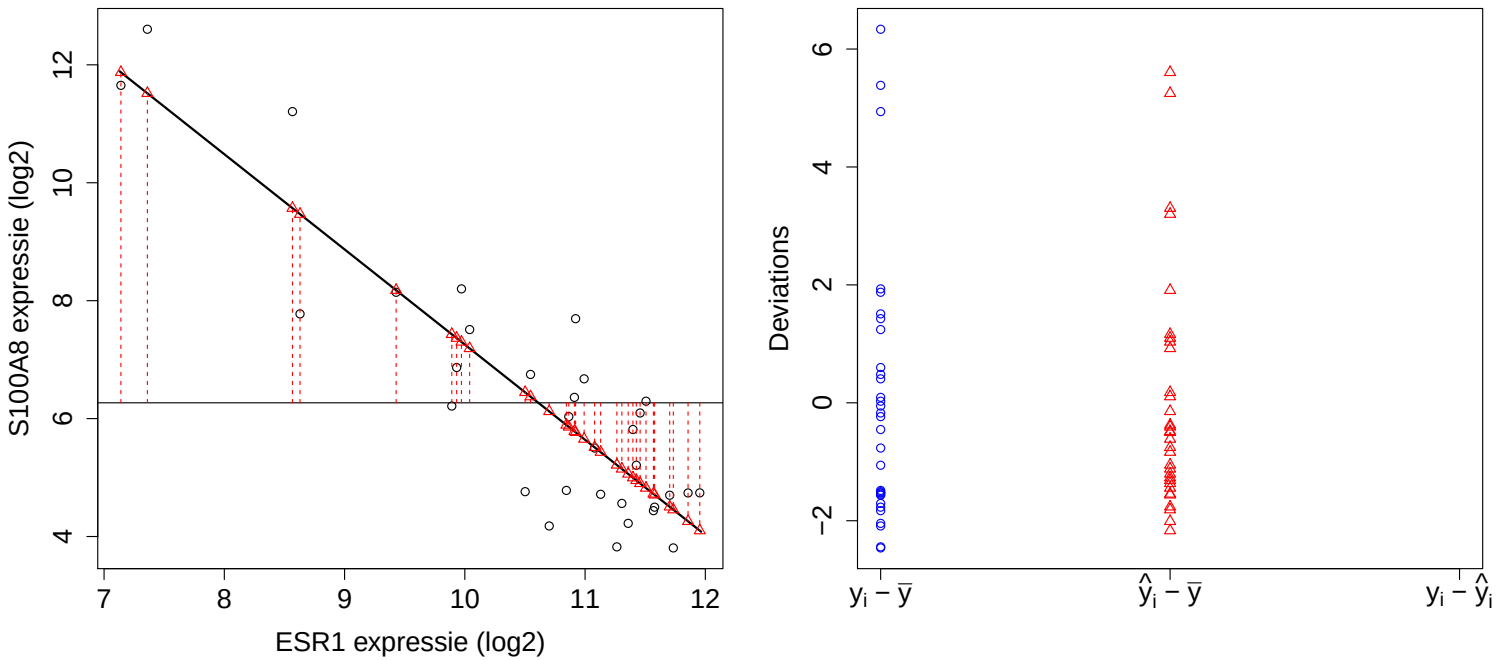
$$\text{SSTot} = \text{SSR} + \text{SSE}.$$

$$\begin{aligned}
\text{SSTot} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\
&= \text{SSE} + \text{SSR}
\end{aligned}$$

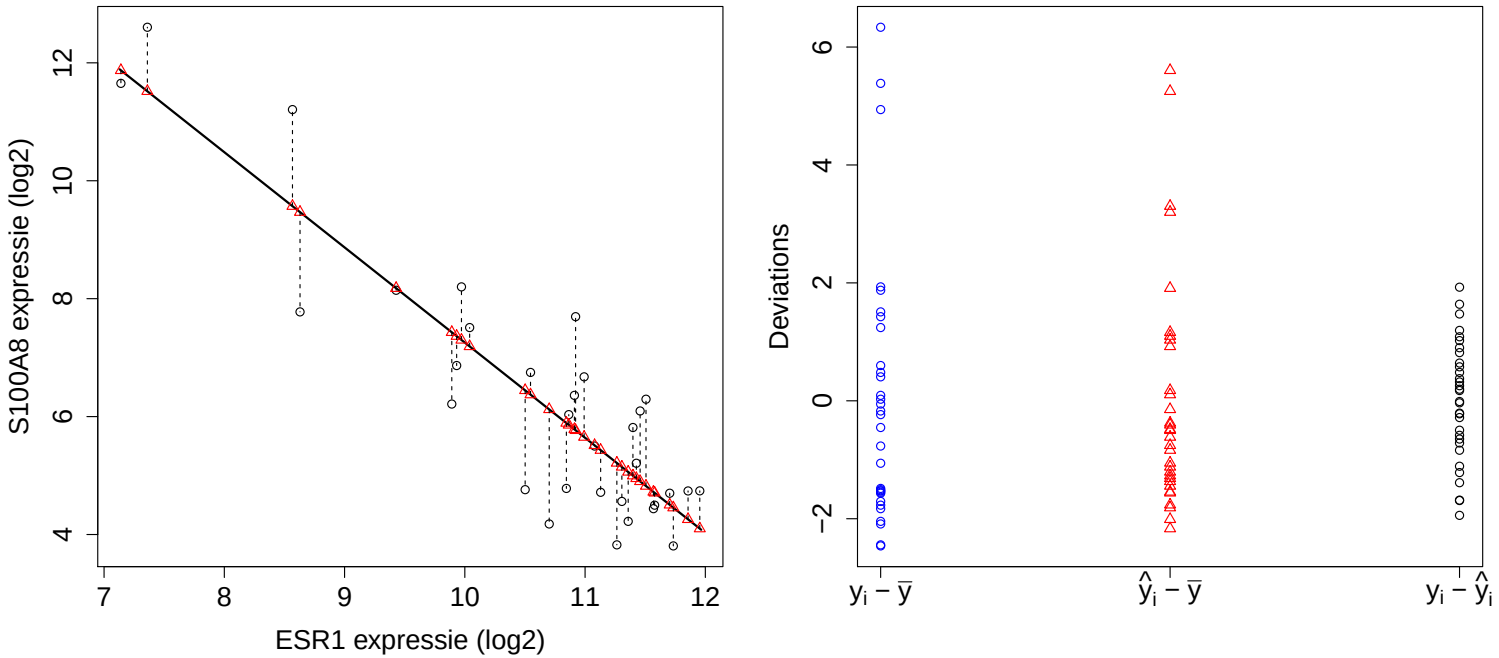
Interpretatie ontbinding totale kwadratensom



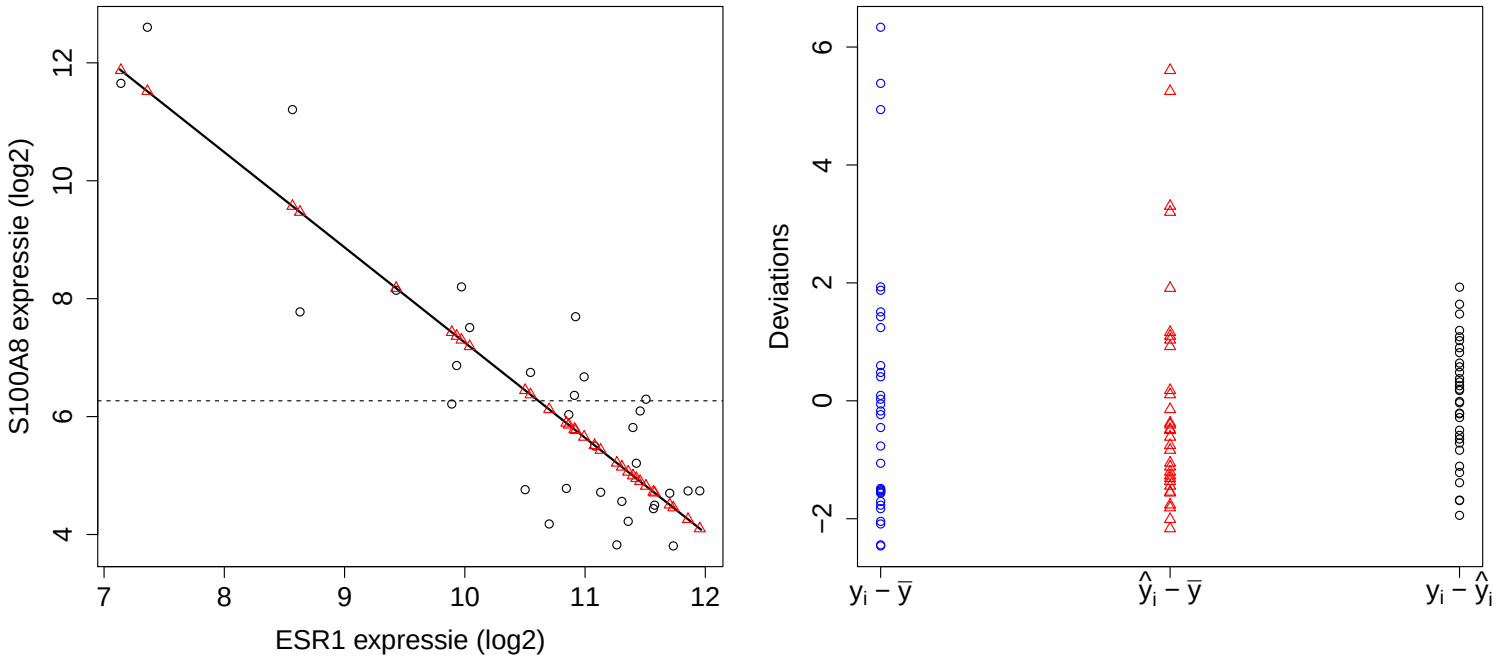
Interpretatie ontbinding totale kwadratensom



Interpretatie ontbinding totale kwadratensom



Interpretatie ontbinding totale kwadratensom



Interpretatie ontbinding totale kwadratensom

De totale variabiliteit in de data (SSTot) wordt gedeeltelijk verklaard door het regressieverband (SSR). De variabiliteit die niet door het regressieverband verklaard wordt, is de residuele variabiliteit (SSE).

Determinatiecoëfficiënt

De determinatiecoëfficiënt wordt gegeven door

$$R^2 = 1 - \frac{\text{SSE}}{\text{SSTot}}.$$

De determinatiecoëfficiënt kan ook geschreven worden als

$$R^2 = 1 - \frac{\text{SSE}}{\text{SSTot}} = 1 - \frac{\text{SSTot} - \text{SSR}}{\text{SSTot}} = \frac{\text{SSR}}{\text{SSTot}}.$$

Het is dus **de fractie van de totale variabiliteit in de steekproef-uitkomsten dat verklaard wordt door het geschatte regressieverband**.

Borstkanker voorbeeld

```
> summary(model3)

Residual standard error: 1.026 on 30 degrees of freedom
Multiple R-squared:  0.7942,Adjusted R-squared:  0.7874
F-statistic: 115.8 on 1 and 30 DF,  p-value: 8.07e-12
```

79.4% van de variabiliteit in de $\log_2$ S100A8 expressie kan worden verklaard door de $\log_2$ ESR1 expressie.

Een grote R^2 is meestal een indicatie dat het model potentieel tot goede predicties kan leiden (kleine SSE), maar de waarde van R^2 is slechts in beperkte mate indicatief voor de p-waarde.
Twee argumenten:

- de p-waarde wordt sterk beïnvloed door SSE, maar niet door SSTot. Ook de steekproefgrootte n heeft een grote invloed op de p-waarde.
- de determinatiecoëfficiënt R^2 wordt door SSE en SSTot bepaald, maar niet door de steekproefgrootte n.

F-Testen

De kwadratensommen vormen de basis van een belangrijke klasse van hypothesetesten.

De F-teststatistiek wordt gedefinieerd als

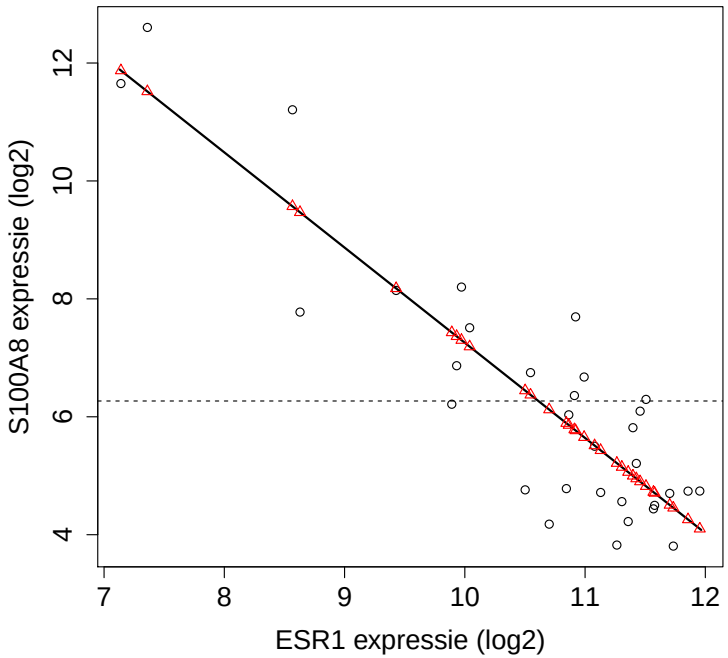
$$F = \frac{\text{MSR}}{\text{MSE}}$$

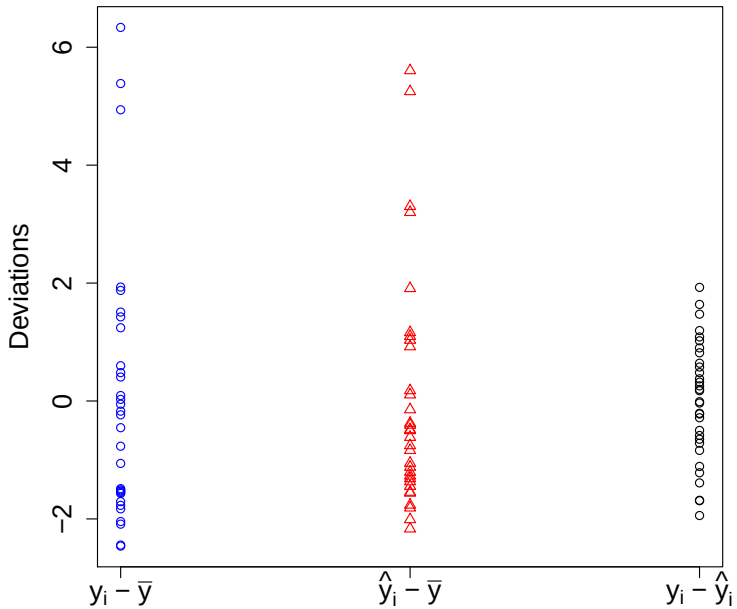
met

$$\text{MSR} = \frac{\text{SSR}}{1} \quad \text{en} \quad \text{MSE} = \frac{\text{SSE}}{n-2}.$$

MSR wordt de **gemiddelde kwadratensom van de regressie** genoemd. De noemers 1 en $n-2$ zijn de vrijheidsgraden van SSR en SSE.

Interpretatie $F = \frac{\text{MSR}}{\text{MSE}}$





F-test in het enkelvoudig lineair regressiemodel

Onder $H_0 : \beta_1 = 0$,

$$F = \frac{\text{MSR}}{\text{MSE}} \overset{H_0}{\sim} F_{1,n-2}.$$

De teststatistiek kan enkel gebruikt worden voor het testen tegenover $H_1 : \beta_1 \neq 0$ (tweezijdig alternatief), waarvoor de p-waarde gegeven wordt door

$$p = \text{P}_0\{F \geq f\} = 1 - F_F(f; 1, n-2).$$

De kritieke waarde op het β_0 significantieniveau is $F_{1,n-2;1-\alpha}$.

Anova Tabel

De kwadratensommen en de F-test worden meestal in een zogenaamde **variantie-analyse tabel** of een **anova tabel** gerapporteerd.

Borstkanker voorbeeld

```
> anova(model3)
Analysis of Variance Table

Response: logS100A8
          Df  Sum Sq Mean Sq F value   Pr(>F)
logESR1    1 121.814 121.814   115.8 8.07e-12 ***
Residuals 30  31.559   1.052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We besluiten dus dat er een extreem significant lineair verband is tussen de log_2 ESR1 expressie en de log_2 S100A8 expressie. De F-test is tweezijdig. Door te kijken naar het teken van $\hat{\beta}_1$ ($\hat{\beta}_1 = -1.615$) kunnen we tevens besluiten dat er een negatieve associatie is tussen beiden.

- Merk op dat de p-waarde van de F-test en de p-waarde van de tweezijdige t-test exact gelijk zijn.
- Voor het enkelvoudig lineair regressie-model zijn beide testen equivalent!

```
> summary(model3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.401      1.603   14.60 3.57e-15 ***
logESR1       -1.615      0.150  -10.76 8.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.026 on 30 degrees of freedom
Multiple R-squared:  0.7942,Adjusted R-squared:  0.7874
F-statistic: 115.8 on 1 and 30 DF,  p-value: 8.07e-12
```

Dummy variabelen

Dummy variabelen

Het lineaire regressiemodel kan ook gebruikt worden voor het vergelijken van twee gemiddelden.

Borstkanker voorbeeld

Is er een verschil in de gemiddelde leeftijd van de patiënten met onaangetaste lymfeknopen en patiënten waarvoor de lymfeknopen werden verwijderd?

We definiëren de dummy variabele

$$x_i = \begin{cases} 1 & \text{aangetaste lymfeknopen} \\ 0 & \text{onaangetaste lymfeknopen} \end{cases}.$$

De groep met $x_i = 0$ wordt de referentiegroep genoemd.
Het regressiemodel blijft ongewijzigd,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

met ε_i i.i.d. $N(0, \sigma^2)$.

Gezien x_i slechts twee waarden kan aannemen, is het eenvoudig het regressiemodel voor beide waarden van x_i afzonderlijk te bekijken:

$$\begin{aligned} Y_i &= \beta_0 + \varepsilon_i \text{onaangetaste lymfeknopen} (x_i = 0) \\ Y_i &= \beta_0 + \beta_1 + \varepsilon_i \text{ aangetaste lymfeknopen} (x_i = 1). \end{aligned}$$

Dus

$$\begin{aligned} \mathsf{E}\{Y_i \mid x_i = 0\} &= \beta_0 \\ \mathsf{E}\{Y_i \mid x_i = 1\} &= \beta_0 + \beta_1, \end{aligned}$$

waaruit direct de interpretatie van β_1 volgt:

$$\beta_1 = \mathsf{E}\{Y_i \mid x_i = 1\} - \mathsf{E}\{Y_i \mid x_i = 0\},$$

i.e. β_1 is het gemiddelde verschil in leeftijd tussen patiënten met aangetaste lymfeknopen en patiënten met onaangetaste lymfeknopen (referentiegroep).

Met de notatie $\mu_1 = \mathsf{E}\{Y_i \mid x_i = 0\}$ en $\mu_2 = \mathsf{E}\{Y_i \mid x_i = 1\}$ wordt dit

$$\beta_1 = \mu_2 - \mu_1.$$

(Noot: de indexen 1 en 2 mogen gerust vervangen worden door 0 en 1 om explicieter naar $x_i = 0$ en $x_1 = 1$ te verwijzen; dan wordt $\beta_1 = \mu_1 - \mu_0$.)
Er kan aangetoond worden dat

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y}_1 \quad \text{(steekproefgemiddelde in referentiegroep)} \\ \hat{\beta}_1 &= \bar{Y}_2 - \bar{Y}_1 \quad \text{(schatter van effectgrootte)} \\ \text{MSE} &= S_p^2. \end{aligned}$$

Output linear model

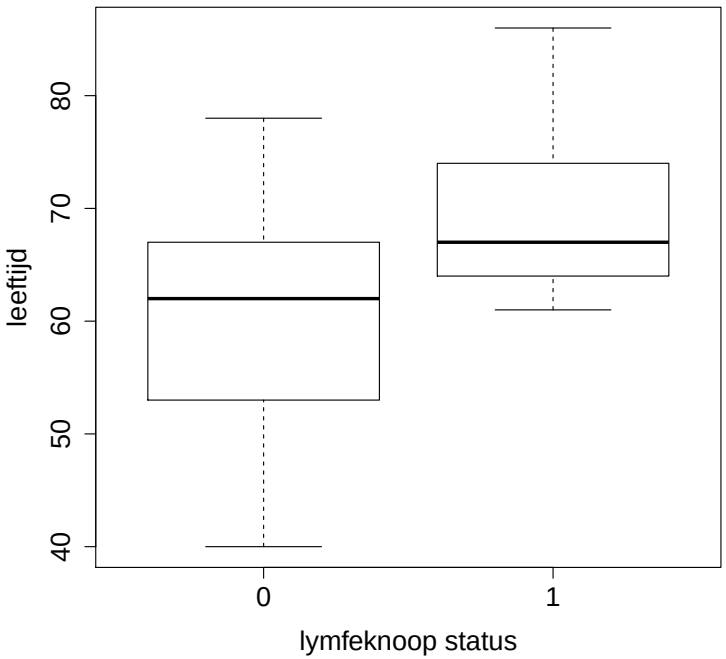
De test voor het testen van $H_0 : \beta_1 = 0$ kan gebruikt worden voor het testen van de nulhypothese van de two-sample t-test, $H_0 : \mu_1 = \mu_2$.

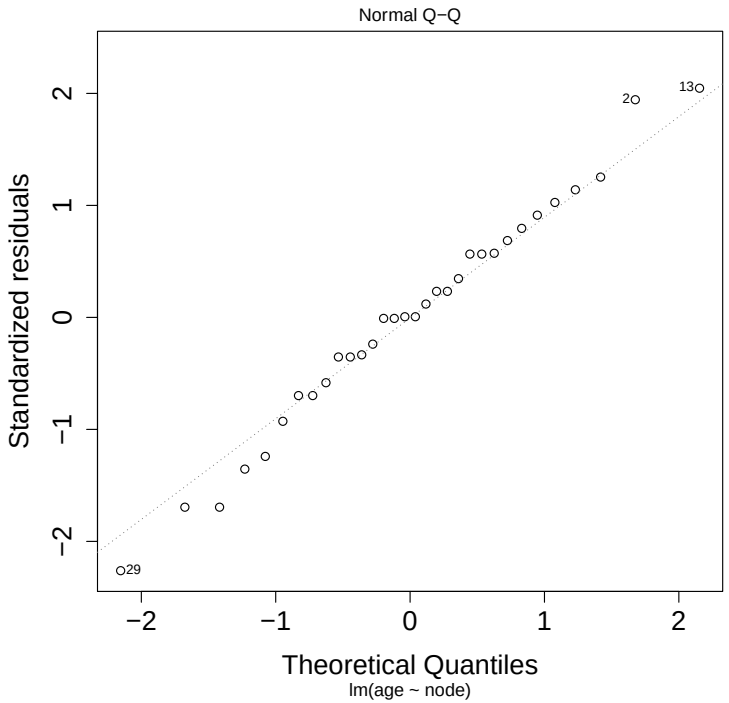
```
> model4 <- lm(age~node,borstkanker)
> summary(model4)

Call:
lm(formula = age ~ node, data = borstkanker)

Residuals:
     Min       1Q   Median       3Q      Max
-19.9474  -5.3269   0.0526   5.3026  18.0526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.947      2.079  28.834  < 2e-16 ***
node           9.130      3.262   2.799  0.00888 **
---
```


leeftijd
40
50
60
70
80
0
1
lymfeknoop status


Normal Q−Q
Standardized residuals
−2
−1
0
1
2
29
2
13
Theoretical Quantiles
lm(age ~ node)

Output two-sample t-test

```
> t.test(age~node,borstkanker,var.equal=TRUE)

Two Sample t-test

data:  age by node
t = -2.7988, df = 30, p-value = 0.008879
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -15.791307  -2.467802
sample estimates:
mean in group 0 mean in group 1
       59.94737        69.07692
```

Observationele studie

- Merk op dat dit een observationele studie is.
- Mogelijks verschillende de patiënten niet enkel in lymfeknoop status maar ook in andere karakteristieken en zijn beide groepen patiënten niet vergelijkbaar!
- We hebben immers niet kunnen randomiseren!
- We kunnen enkel besluiten dat er een associatie is tussen de lymfeknoop status en de leeftijd. Het is dus niet noodzakelijkerwijs een causaal verband!
- Het is steeds moeilijk om causale verbanden te trekken op basis van observationele studies gezien confounding kan optreden.