Lineaire regressie

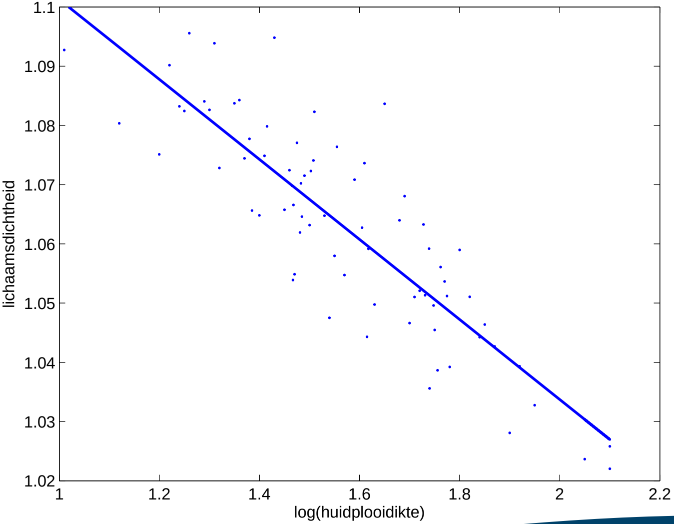
Sandra Van Aert

8 december 2011

Voorbeeld

- meten van de lichaamsdichtheid (gewicht van het lichaam per volume-eenheid)
- grote hoeveelheid vet betekent lage lichaamsdichtheid
- lichaamsdichtheid is moeilijk rechtstreeks te meten
- huidplooidikte is wel eenvoudig te meten
- bestaat er een verband tussen lichaamsdichtheid en huidplooidikte?

Lineair verband lichaamsdichtheid en log(huidplooidikte)



Lineair verband lichaamsdichtheid en log(huidplooidikte)

- Vergelijking regressierechte:
  lichaamsdichtheid=
  1.1688-0.0676log(huidplooidikte)
- toename van log(huidplooidikte) met 1 geeft een verlaging van lichaamsdichtheid met 0.0676 g/cm^3

Lineaire regressie

- bestuderen van de relatie tussen 2 variabelen
- men veronderstelt een lineair verband:

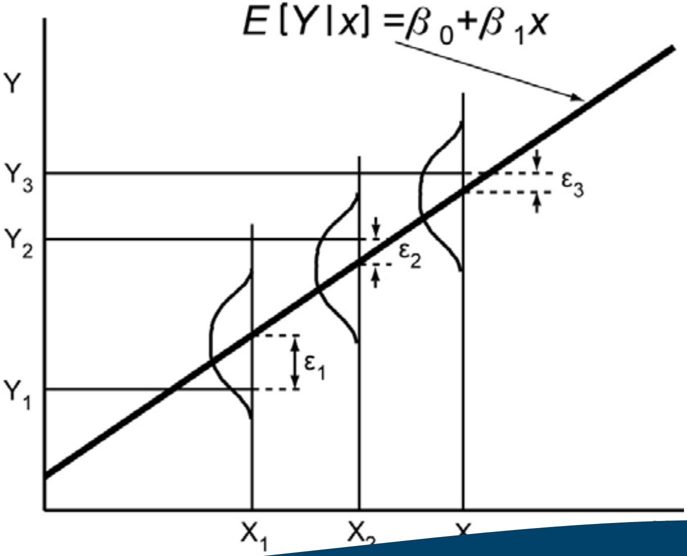
$$Y = \beta_0 + \beta_1 x + \epsilon$$

- Y random variabele = respons of afhankelijke variabele
- x niet-random variabele = verklarende of onafhankelijke variabele
- regressie coëfficiënten:
  β_0 de intercept
  β_1 de helling

Lineaire regressie

- ϵ storingsterm
- verklaart de residuele variatie in Y
- veronderstelling:
  onafhankelijk en normaal verdeelde residu's
  en constante variantie
  $E(\epsilon) = 0$
  $var(\epsilon) = \sigma^2$
- Hieruit volgt:
  $E(Y|x) = \beta_0 + \beta_1 x$
  $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

Grafische voorstelling regressiemodel

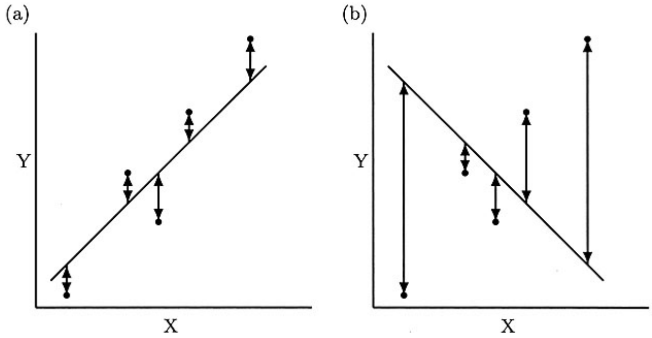


Hoe?

Kleinste kwadraten methode

- minimaliseren kleinste kwadraten criterium $S^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$
- levert rechte met minimale spreiding
- $Y = \hat{\beta}_0 + \hat{\beta}_1 x$

Hoe?



rechte met (a) kleine en (b) grote spreiding

Kleinste kwadraten schatters

$$
\begin{aligned}
\frac{\partial S^2}{\partial \beta_0} &= -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \\
\frac{\partial S^2}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i\left(y_i - \beta_0 - \beta_1 x_i\right) = 0
\end{aligned}
$$

$$
\begin{aligned}
n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\
\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i
\end{aligned}
$$

Kleinste kwadraten schatters

Uit

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

volgt

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \overline{y} - \hat{\beta}_1 \overline{x} \end{aligned}$$

Kleinste kwadraten schatters

Invullen in

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

geeft

$$\left(\overline{y} - \hat{\beta}_1 \overline{x}\right) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

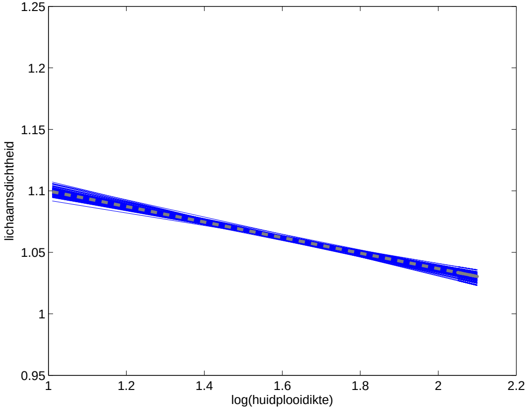
$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}
\end{aligned}$$

Kleinste kwadraten schatters

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

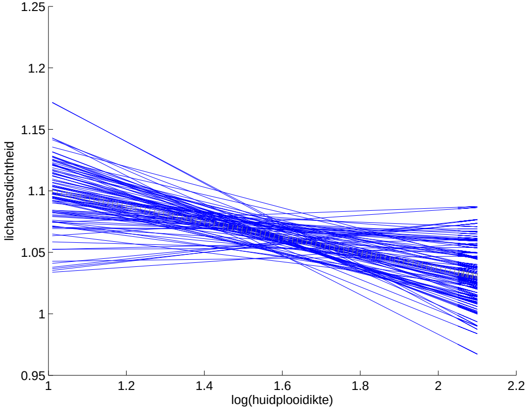
$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} \\
&= \frac{s_{XY}}{s_X^2} \\
&= r_{XY} \frac{s_Y}{s_X}
\end{aligned}$$

Eigenschappen



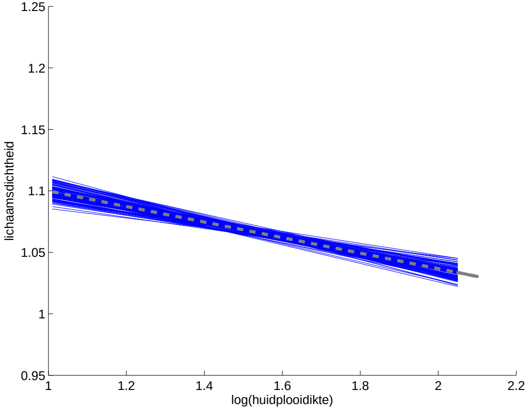
kleinste kwadraten rechten bekomen uit 100 steekproeven ($\sigma = 0.00854$ en $n = 74$)

Eigenschappen



kleinste kwadraten rechten bekomen uit 100 steekproeven ($\sigma = 10 \times 0.00854$ en $n = 74$)

Eigenschappen



kleinste kwadraten rechten bekomen uit 100 steekproeven ($\sigma = 0.00854$ en $n = 13$)

Eigenschappen

$$
\begin{aligned}
var\left(\hat{\beta}_0\right) &= \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{(n-1)s_X^2}\right)\\
var\left(\hat{\beta}_1\right) &= \frac{\sigma^2}{(n-1)s_X^2}\\
cov\left(\hat{\beta}_0,\hat{\beta}_1\right) &= -\frac{\overline{x}\sigma^2}{(n-1)s_X^2}
\end{aligned}
$$

Variantie op de schattingen wordt kleiner wanneer:

- de variantie σ^2 kleiner wordt
- steekproefgrootte n groter wordt
- de waarden x_i gelijkmatiger verspreid liggen

Eigenschappen

Meest precieze lineaire onvertekende schatter of Best Linear Unbiased Estimator BLUE of Minimum Variance Unbiased Estimator MVUE

- minimale variantie
- zuivere of onvertekende schatters:

$$
\begin{aligned}
E\left(\hat{\beta}_0\right) &= \beta_0 \\
E\left(\hat{\beta}_1\right) &= \beta_1
\end{aligned}
$$

Eigenschappen

Verdeling van de schattingen:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_X^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)s_X^2}\right)$$

Onvertekende schatter voor $\sigma^2 = var(\epsilon_i) = E\left(\epsilon_i^2\right)$:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2$$

met $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Betrouwbaarheidsintervallen

Uitgaande van de verdelingsfuncties van de schattingen volgt:

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{(n-1)s_X^2}}} \sim N(0,1)$$

Wanneer σ^2 ongekend is, volgt:

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}}} \sim t_{n-2}$$

Betrouwbaarheidsintervallen

$$P\left(-t_{\alpha/2;n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}}} \leq t_{\alpha/2;n-2}\right) = 1 - \alpha.$$

$100(1-\alpha)\%$ betrouwbaarheidsinterval voor β_1:

$$P\left(\hat{\beta}_1 - t_{\alpha/2;n-2}\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2;n-2}\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}}\right)$$

Hypothesetoets

Kan de respons variabele verklaard worden door x?

$$H_0 : \beta_1 = 0 \text{ en } H_a : \beta_1 \neq 0$$

- Verwerp H_0 indien

$$\hat{\beta}_1 < -t_{\alpha/2;n-2}\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}} \text{ of } \hat{\beta}_1 > +t_{\alpha/2;n-2}\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}}$$

- Aanvaard H_0 indien

$$-t_{\alpha/2;n-2}\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}} \leq \hat{\beta}_1 \leq +t_{\alpha/2;n-2}\frac{\hat{\sigma}}{\sqrt{(n-1)s_X^2}}$$

Puntschatting

Voorspelling maken voor Y gegeven $x = x_0$

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

met $\hat{y}_0$ de geschatte respons waarde

- dit is slechts een schatting...
- hoe rekening houden met onzekerheid in gefitte regressierechte?
- hoe rekening houden met spreiding rondom de gefitte rechte?

constructie van betrouwbaarheidsintervallen

Betrouwbaarheidsinterval voor de gemiddelde respons

Variantie op de gemiddelde respons $\hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\begin{aligned} var\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) &= var\left(\hat{\beta}_0\right) + x_0^2 var\left(\hat{\beta}_1\right) + 2x_0 cov\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{(n-1)s_X^2}\right) \end{aligned}$$

waarbij σ^2 geschat kan worden door $\hat{\sigma}^2$

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{(n-1)s_X^2}}} \sim t_{n-2}$$

Betrouwbaarheidsinterval voor de gemiddelde respons

$100(1-\alpha)\%$ betrouwbaarheidsinterval voor $\beta_0 + \beta_1 x_0$

$$P\left(\hat{L} \leq \beta_0 + \beta_1 x_0 \leq \hat{U}\right) = 1 - \alpha$$

met

$$\hat{L} = \hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{\alpha/2;n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{(n-1)s_X^2}}$$

$$\hat{U} = \hat{\beta}_0 + \hat{\beta}_1 x + t_{\alpha/2;n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{(n-1)s_X^2}}$$

hangt af van $x_0 - \overline{x}$

Predictie-interval voor de respons

Variantie op de respons $\hat{y}_m = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\begin{aligned} var\left(Y_m - \hat{y}_m\right) &= var\left(Y_m - (\hat{\beta}_0 + \hat{\beta}_1 x_m)\right) \\ &= var\left(Y_m\right) + var\left(\hat{\beta}_0 + \hat{\beta}_1 x_m\right) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{(n-1)s_X^2}\right) \end{aligned}$$

waarbij σ^2 geschat kan worden door $\hat{\sigma}^2$.
Verder geldt:

$$E(Y_m - \hat{y}_m) = 0$$

Predictie-interval voor de respons

$$\frac{Y_m - \hat{y}_m}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_X^2}}} \sim t_{n-2}$$

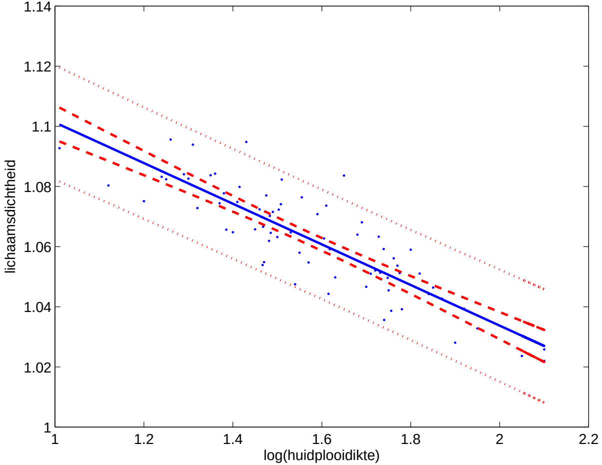
$100(1-\alpha)\%$ predictie-interval voor Y_m

$$P\left(\hat{L} \leq Y_m \leq \hat{U}\right) = 1 - \alpha$$

$$\hat{L} = \hat{y}_m - t_{\alpha/2;n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_X^2}}$$

$$\hat{U} = \hat{y}_m + t_{\alpha/2;n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_X^2}}$$

Betrouwbaarheids- en predictie-interval



Nagaan van modelonderstellingen

- lineariteit van de respons variabelen in de parameters β_0 en β_1
  puntenwolk met bijbehorende regressierechte
- normaal verdeelde residu's ϵ met gemiddelde 0 en identieke variantie σ^2 voor alle $i = 1, \ldots, n$ ofwel $\epsilon_i \sim N(0, \sigma^2)$
  kwantieldiagram of Shapiro-Wilk toets op basis van de residu's
- residu's ϵ_i onafhankelijk van elkaar
  residuplot (residu's r_i uitgezet tegenover x_i)

Vragen over theorie of oefeningen ???

- Gelieve al uw vragen door te sturen vóór 18 december!
- Alle vragen zullen overlopen worden tijdens de laatste les op dinsdag 20 december.