Stochastiek 2

Inleiding in de Mathematische Statistiek

Lineaire regressie

Tot nu toe

Setting: meestal $X_1, \ldots, X_n$ o.o. $X_i \sim p_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$, $k = 1, 2$.

- Methoden voor schatten van θ (ad hoc, MLE, Bayes, momenten)
- Methoden voor toetsen van hypotheses (Gauss-toets, binomiale toets, t-toetsen, likelihood-ratiotoets)
- Methoden voor betrouwbaarheidsgebieden ((bijna) pivots, likelihood-ratiogebieden, credible sets)

Lineaire modellen

Geïnteresseerd in verband tussen verschillende variabelen.

Data: $(\underbrace{Y_i}_{\text{response}}, \underbrace{x_{i,1}, \ldots, x_{i,p}}_{\text{“verklarend”}})$, $i = 1, \ldots, n$.

Model: $Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + e_i$, met e_i i.i.d., verwachting 0.

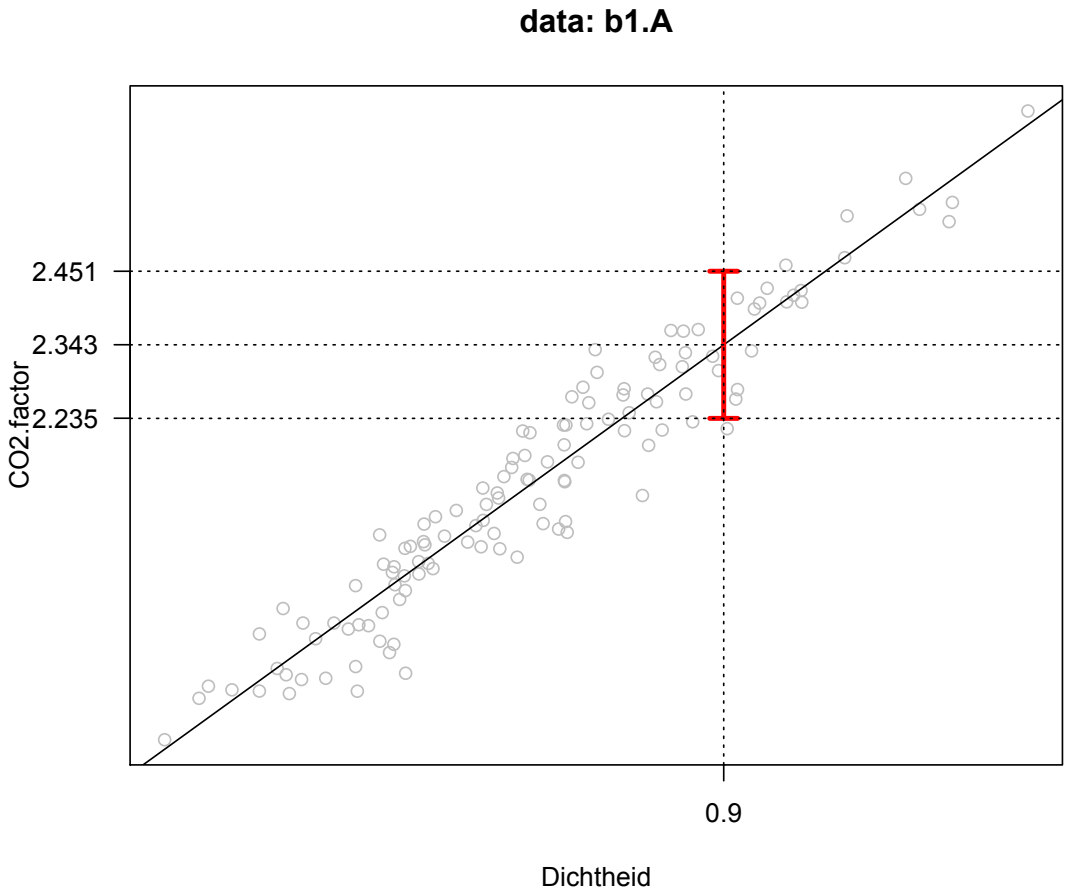
Vragen:

- Schat de parameter $\beta = (\beta_1, \ldots, \beta_p)$.
- Is β_i significant $\neq 0$?
- Wat is een betrouwbaarheidsinterval voor β_i?
- Welke verklarende variabelen moeten er in het model opgenomen worden?
- ...

Voorbeeld 1: CO_2 emissie

Y_i: CO_2 uitstoot

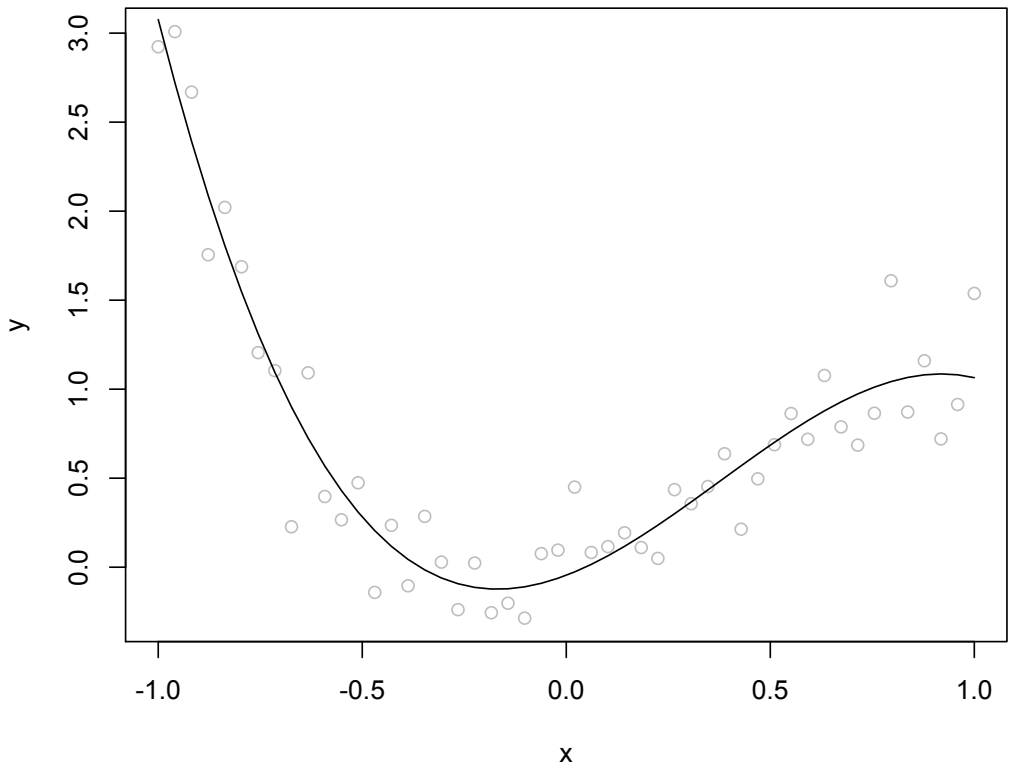
x_i: dichtheid van brandstof



Voorbeeld 2: fitten van 3e-graads polynoom

Data: (x_i, Y_i), $i = 1, \ldots, 50$.

Model: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i$.



Enkelvoudige lineaire regressie

Enkelvoudige lineaire regressiemodel

Data: (x_i, Y_i), $i = 1, \ldots, n$.

Model:

$$Y_i = \alpha + \beta x_i + e_i, \qquad , i = 1, \ldots, n,$$

met $e_1, \ldots, e_n$ o.o. en $N(0, \sigma^2)$-verdeeld.

Parameter:

$$\theta = (\alpha, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+.$$

Enkelvoudige lineaire regressiemodel - schatten

MLS $(\hat{\alpha}, \hat{\beta})$ voor (α, β): punt waar

$$(\alpha, \beta) \mapsto \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2$$

minimaal is. Dit zijn de kleinste kwadratenschatters.

MLS $\hat{\sigma}^2$ voor σ^2:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

(Zuivere schatter voor σ^2:

$$\tilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.)$$

Enkelvoudige lineaire regressiemodel - residuen, R^2

Residuen: $Y_i - \hat{\alpha} - \hat{\beta} x_i$

Residuele kwadraatsom: $SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$

Totale kwadraatsom: $SS_{tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

R^2 - determinatiecoëfficient: fractie van de variantie verklaard door regressie op the x_i

$$r^2_{x,Y} = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^2}.$$

Altijd $r^2_{x,Y} \in [0, 1]$. Als $r^2_{x,Y} \approx 1$: punten dichtbij rechte lijn, als $r^2_{x,Y} \approx 0$: punten sterk verspreid rond rechte lijn, of geen lineair verband.

$\rightarrow$ chirps in R

Enkelvoudige lineaire regressiemodel - toetsen

Interessante parameter: β. Verwerpen van $H_0 : \beta = 0$ wil zeggen "er is een significant verband tussen de x_i en de Y_i".

Voor gegeven $\beta_0 \in \mathbb{R}$, beschouw de hypothese $H_0 : \beta = \beta_0$.
Idee: gebruik een toetsingsgrootheid gebaseerd op the MLS $\hat{\beta}$.

We weten (Opgave 7.4!): $\hat{\beta} \sim N(\beta, \sigma^2/(n-1)s_x^2)$. Een (zuivere) schatter voor de variantie is

$$\widehat{var}\,\hat{\beta} = \frac{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Enkelvoudige lineaire regressiemodel - toetsen

Gebruiken als toetsingsgrootheid:

$$T = \frac{\hat{\beta} - \beta_0}{\sqrt{\widehat{var}\,\hat{\beta}}}.$$

In sectie 7.2.2.3 bewijzen we: onder $H_0 : \beta = \beta_0$ geldt $T \sim t_{n-2}$.

Conclusie: De toets "verwerp $H_0 : \beta = \beta_0$ als $|T| \geq t_{n-2,1-\alpha_0/2}$ is een toets van niveau α_0.

(Alternatief: likelihood-ratiotoets)

$\rightarrow$ chirps in R

Enkelvoudige lineaire regressiemodel - bi

Willen een betrouwbaarheidsinterval voor de parameter β.

Pivot: $T = (\hat{\beta} - \beta)/\sqrt{\widehat{var}\,\hat{\beta}}$.
We weten (sectie 7.2.2) dat $T \sim t_{n-2}$.

Er geldt dus dat

$$P_{\alpha,\beta,\sigma^2}(-t_{n-2,1-\alpha_0/2} \leq T \leq t_{n-2,1-\alpha_0/2}) = 1 - \alpha_0.$$

Conclusie: een α_0-b.i. voor β is

$$\beta = \hat{\beta} \pm t_{n-2,1-\alpha_0/2}\sqrt{\widehat{var}\,\hat{\beta}}.$$

$\rightarrow$ chirps in R