Stochastiek 2

Inleiding in de Mathematische Statistiek

Meervoudige 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:

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

Matrix notatie:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Meervoudige lineaire regressie

Data: $Y = (Y_1, \ldots, Y_n)^T$.

Model:

$$Y = X\beta + e,$$

met X een (deterministische) $n \times p$ matrix (de design matrix), $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ een onbekende parameter vector en $e = (e_1, \ldots, e_n)$, e_i o.o. en $N(0, \sigma^2)$-verdeeld.

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

Als $X = (x_{ij})$, dan voor iedere i:

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i.$$

Dummy-variabelen

Categorische verklarende variabele: variabele die waarden aanneemt in een verzameling van een eindig aantal klassen. (Bijv: man/vrouw, maand, ...). (Engels: factors, met eindig aantal levels.)

Categorische variabelen kunnen worden opgenomen in een regressiemodel mbv dummy-variabelen: $\{0, 1\}$-variabelen of $\{-1, 1\}$-variabelen die aangeven tot welke klasse een categorische variabele behoort.

Voorbeeld: lengte data van mannen en vrouwen. Y_i: lengte van persoon i,

$$x_i = \begin{cases} 0, & \text{als persoon } i \text{ een vrouw is,} \\ 1, & \text{als persoon } i \text{ een man is.} \end{cases}$$

Model: $Y_i = \alpha + \beta x_i + e_i$.

Maximum likelihood-schatter

Stel dat X volle rang heeft. (Anders X aanpassen.)

MLS $\hat{\beta}$ voor β: punt waar

$$\beta \mapsto \|Y - X\beta\|^2$$

minimaal is. Er geldt: $\hat{\beta} = (X^T X)^{-1} X^T Y$.

MLS $\hat{\sigma}^2$ voor σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\|Y - X\hat{\beta}\|^2.$$

Gefitte waarden, residuen, R^2

Gefitte waarden: $X\hat{\beta}$: orth. projectie van Y op kolomruimte X.

Residuen: $Y - X\hat{\beta}$: orth. proj. van Y op orth. compl. kolomr. X

$\Rightarrow$ gefitte waarden en residuen zijn ongecorreleerd. (plots)

Als voorheen: residuele en totale kwadraatsommen

$$SS_{res} = \|Y - X\hat{\beta}\|^2, \qquad SS_{tot} = \|Y - \bar{Y}1\|^2.$$

Determinatiecoëfficient, of R^2,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

$\rightarrow$ cement voorbeeld

Intermezzo: verdeling MLS

Stelling.

Stel, X heeft volle rang p.

(i) Voor alle j, $\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj})$.

(ii)

$$\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

(iii) De MLS $\hat{\beta}$ en de residuen $Y - X\hat{\beta}$ zijn onafhankelijk.

Schets van bewijs.

(i) en (iii): volgen uit de expliciete uitdrukking voor $\hat{\beta}$, algebra en normaliteit. (ii): merk op dat $Y - X\hat{\beta} = Pe$, met P de orthogonale projectie op het orthogonaal complement van de kolomruimte van X. □

Intermezzo: zuivere schatter voor σ^2

Gevolg.

Stel dat X volle rang p heeft. Dan is

$$S^2 = \frac{1}{n-p}\|Y - X\hat{\beta}\|^2$$

een zuivere schatter voor σ^2.

Toetsen - 1

Voor vast j, beschouw de hypotheses $H_0 : \beta_j = 0$ tegen $H_1 : \beta_j \neq 0$. Idee: gebruik een toetsingsgrootheid gebaseerd op $\hat{\beta}_j$.

We weten dat $\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj})$ en dat S^2 een zuivere schatter is voor σ^2.

We gebruiken als toetsingsgrootheid

$$T = \frac{\hat{\beta}_j}{\sqrt{S^2 (X^T X)^{-1}_{jj}}}.$$

(NB: (7.4) in het boek is fout!!)

Toetsen - 2

We hebben

$$T = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}} \Big/ \sqrt{\frac{(n-p)S^2/\sigma^2}{n-p}}.$$

Vanwege de stelling, $T \sim t_{n-p}$ onder H_0. Dus toets voor $H_0 : \beta_j = 0$ van niveau α_0:

"verwerp H_0 voor $|T| \geq t_{n-p,1-\alpha_0/2}$".

$\rightarrow$ cement voorbeeld

Meer vragen

(i) lineare modellen voor onderzoeken van invloed van categorische variabelen (sectie 7.3, ANOVA)

(ii) toetsen van meer gecompliceerde hypothesen (sectie 7.2.2.3, sectie 7.3.2)

(iii) verifieren van modelaannames

(iv) niet-lineaire modellen (sectie 7.4)

(v) discrete ipv continue repsonse: classificatie (sectie 7.5)

(vi) modelselectie (hoofdstuk 8)

(vii) ...