Zum linearen Modell, klassischer Kleinste-Quadrate- und Maximum-Likelihood-Schätzung

1. Vortrag des Seminars Erweiterungen des linearen Regressionsmodells und genomische Anwendungen in der Biomedizin, WS 2014/2015

Matthias Birkner

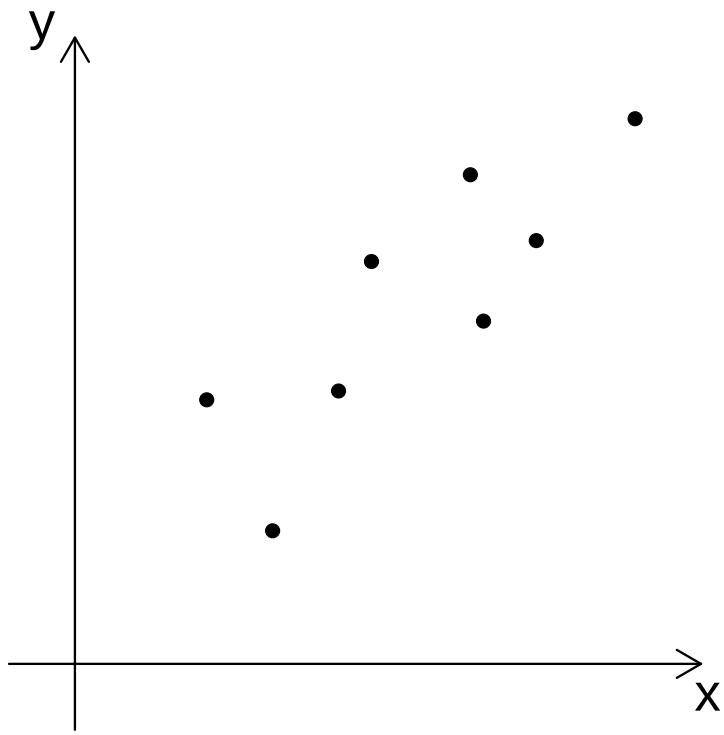
3.11.2014

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Inhalt

Erinnerung: Lineare Regression

Daten:
n Wertepaare (x_i, y_i)



Erinnerung: Lineare Regression

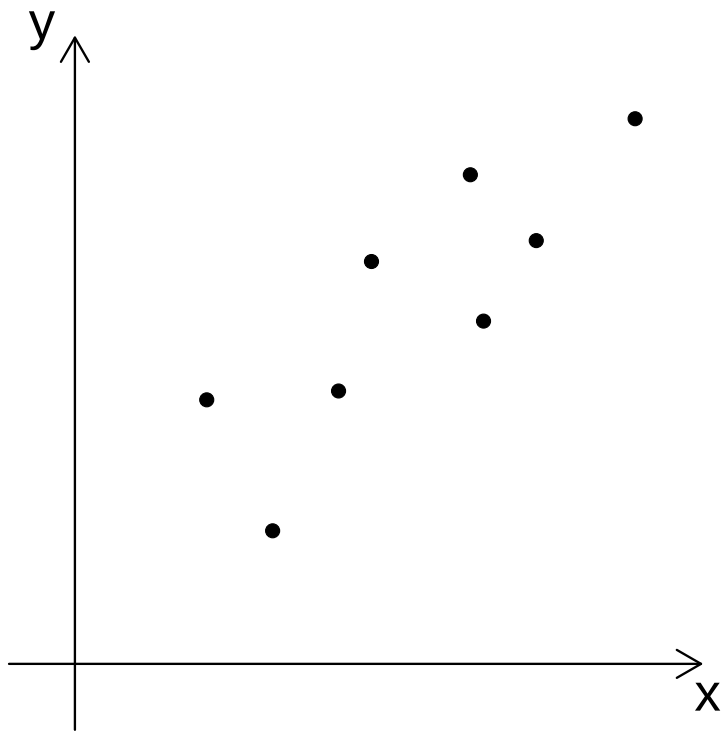
Daten:
n Wertepaare (x_i, y_i)

Modell: (approximativ) linearer Zusammenhang

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

mit ε_i u.a., $\mathbb{E}[\varepsilon_i] = 0$,
$\mathrm{Var}[\varepsilon_i] = \sigma^2$

(hier: $p = 2$ Parameter β_0, β_1)



Erinnerung: Lineare Regression

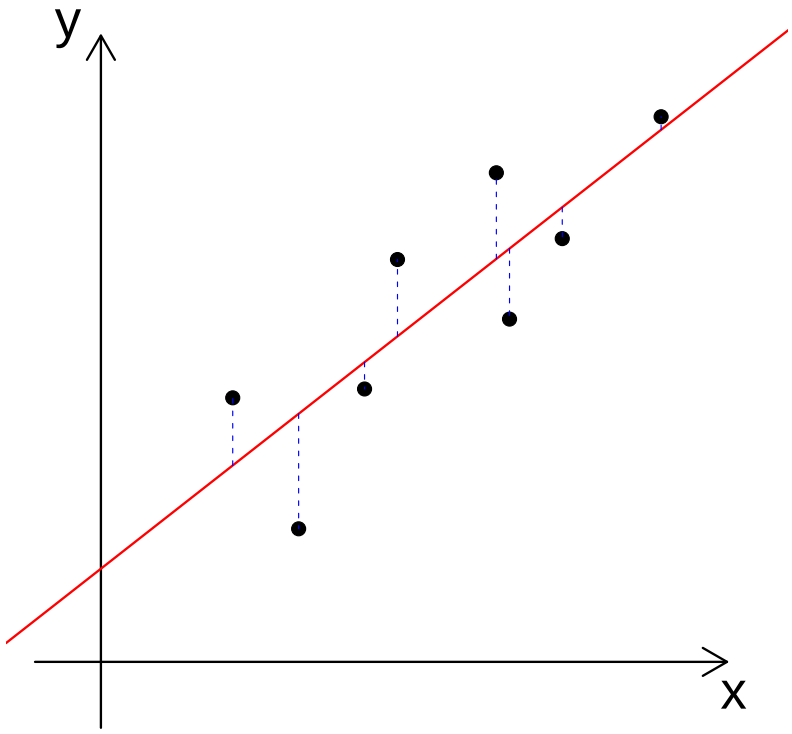
Daten:
n Wertepaare (x_i, y_i)

Modell: (approximativ) linearer Zusammenhang

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

mit ε_i u.a., $\mathbb{E}[\varepsilon_i] = 0$,
$\mathrm{Var}[\varepsilon_i] = \sigma^2$

(hier: $p = 2$ Parameter β_0, β_1)



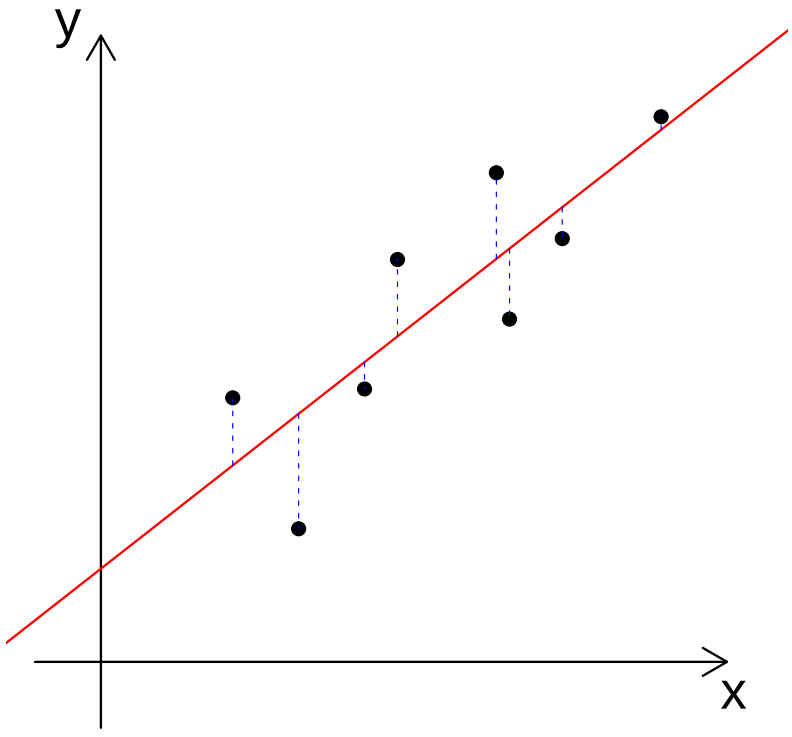
Kleinste-Quadrate-Schätzer: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ derart, dass
$\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \stackrel{!}{=}$ minimal

Erinnerung: Lineare Regression

Daten:
n Wertepaare (x_i, y_i)

Modell: (approximativ) linearer Zusammenhang

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



mit ε_i u.a., $\mathbb{E}[\varepsilon_i] = 0$,
$\mathrm{Var}[\varepsilon_i] = \sigma^2$

(hier: $p = 2$ Parameter β_0, β_1)

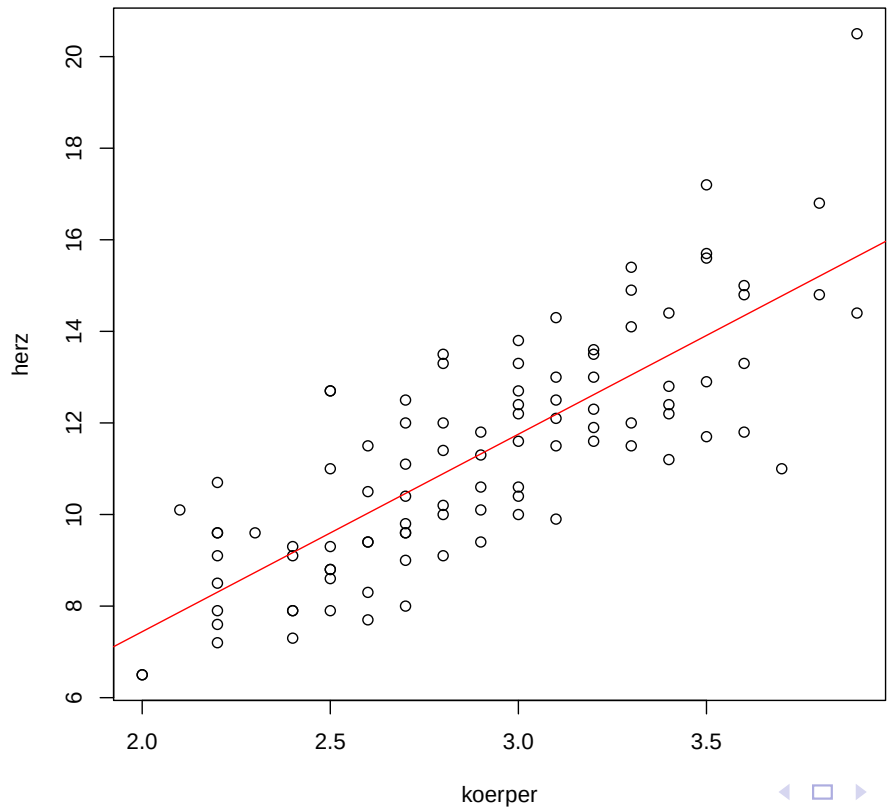
Kleinste-Quadrate-Schätzer: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ derart, dass
$\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \stackrel{!}{=}$ minimal

(mit Lösung $\widehat{\beta}_1 = \mathrm{cov}(x, y)/\sigma_x^2$, $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$,
wo $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$,
$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$)

Lineare Regression mit R

Rs cats-Datensatz, aus Fisher, Biometrics, 3, 65-68 (1947)
(Wir betrachten nur die 97 Kater):

```
> data(cats, package="MASS"); attach(cats)
> koerper <- Bwt[Sex=="M"]; herz <- Hwt[Sex=="M"]
```



Lineare Regression mit R

```
> modell <- lm(herz ~ koerper)
> summary(modell)

Call:
lm(formula = herz ~ koerper)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7728 -1.0478 -0.2976  0.9835  4.8646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1841     0.9983  -1.186    0.239
koerper       4.3127     0.3399  12.688   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

Inhalt

Modellvorstellung

In einer Population besitzt jedes Individuum/Objekt einen (reellen) y-Wert („Zielgröße“) und p (reelle) x-Werte $x_1, \ldots, x_p$ (p „erklärende Variablen“).

Seien $\big((X_1, \ldots, X_p), Y\big)$ die Merkmale eines (zufällig) aus der Population gezogenen Individuums.

Modellvorstellung

In einer Population besitzt jedes Individuum/Objekt einen (reellen) y-Wert („Zielgröße“) und p (reelle) x-Werte $x_1, \ldots, x_p$ (p „erklärende Variablen“).

Seien $\big((X_1, \ldots, X_p), Y\big)$ die Merkmale eines (zufällig) aus der Population gezogenen Individuums.

Wir nehmen an, dass für die Verteilung von Y, gegeben beobachtete (oder je nach Situation auch von uns vorgegebene) Werte $(X_1, \ldots, X_p) = (x_1, \ldots, x_p)$ gilt

$$\mathbb{E}_\beta[Y \,|\, (X_1, \ldots, X_p) = (x_1, \ldots, x_p)] = \beta_1 x_1 + \cdots + \beta_p x_p$$

und

$$\mathrm{Var}_\beta[Y \,|\, (X_1, \ldots, X_p) = (x_1, \ldots, x_p)] = \sigma^2$$

für einen Parametervektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ und ein $\sigma^2 \geq 0$.

Bemerkung:

$$\mathbb{E}_\beta[Y \mid (X_1, \ldots, X_p) = (x_1, \ldots, x_p)] = \beta_1 x_1 + \cdots + \beta_p x_p$$

Der Einfachheit der Formulierung halber integrieren wir hier den Parameter für einen konstanten Wert („Achsenabschnitt“) in das Modell, indem wir annehmen, dass für eine der erklärenden Variablen $\equiv 1$ erfüllt.

Modellvorstellung, 2

Gegeben seien n Beobachtungen, zur i-ten Beobachtung gehört y_i (beobachteter Wert der Zielgröße) und $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ (Werte der p erklärenden Variablen),

wir interpretieren diese als n unabhängige Realisierungen von $\big((X_1, \ldots, X_p), Y\big)$ und schreiben

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

mit ε_i (Realisierungen von) unabhängige(n) Zufallsvariablen mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2$,
wobei wir das „wahre“ β und das „wahre“ σ^2 nicht kennen.

(Je nach Design des Experiments/der Studie stellen wir uns ggfs. vor, dass die Objekte/Individuen bedingt auf gewisse x-Werte ausgewählt wurden.)

Lineares Modell in Matrix-Schreibweise

n Beobachtungen, zur i-ten Beobachtung gehört y_i und $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ (und wir denken im Moment an $n > p$),

$$y_i = \sum_{j=1}^{p} x_{i,j}\beta_j + \varepsilon_i, \quad i = 1, \ldots, p,$$

zusammengefasst

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

mit

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$

$\boldsymbol{X}$ ist die „Designmatrix“ (i-te Zeile gehört zur i-ten Beobachtung)

Kleinste-Quadrate-Ansatz

$n \geq p$ (wir denken an $n \gg p$)

Gegeben beobachtetes $\boldsymbol{y}$ und bekannte/beobachtete Matrix $\boldsymbol{X}$

mit $n \geq p$ (wir denken an $n \gg p$)

Typischerweise besitzt das lineare Gleichungssystem (aufgefasst als Bestimmungsgleichungen für die β_i) $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ keine Lösung.

Kleinste-Quadrate-Schätzer ("least squares")

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} := \operatorname{argmin}\left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \right\}$$

Berechnung (wenigstens prinzipiell)

$\boldsymbol{X}$ habe (maximalen) Rang p, d.h. die p Spalten von $\boldsymbol{X}$ sind linear unabhängig, dann gilt:

Die $p \times p$-Matrix $\boldsymbol{X}^T\boldsymbol{X}$ hat vollen Rang p, ist insbesondere invertierbar, es ist

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

[Falls $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}$ für ein $\boldsymbol{c} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}$, so wäre $\|\boldsymbol{X}\boldsymbol{c}\|_2^2 = \boldsymbol{c}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} = 0$ im Widerspruch zur Ann., dass $\boldsymbol{X}$ Rang p hat.]

Argument (analytisch)

Die „Zielfunktion“

$$\boldsymbol{\beta} \mapsto \sum_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} x_{i,j}\beta_j \right)^2$$

ist quadratisch, das Gleichungssystem

$$\frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} x_{i,j}\beta_j \right)^2 = -2 \sum_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} x_{i,j}\beta_j \right) x_{i,k} \stackrel{!}{=} 0$$

für $k = 1, \ldots, p$ ist in Matrixform geschrieben äquivalent zu

$$\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \qquad \text{„Normalgleichungen“}$$

[Bem.: Diese bestimmen die Lösungsmenge $\{\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\}$ auch in dem Fall, dass $\boldsymbol{X}^T\boldsymbol{X}$ nicht invertierbar ist.]

Argument (geometrisch)

Sei $L = \{\boldsymbol{X}\boldsymbol{\eta} : \boldsymbol{\eta} \in \mathbb{R}^p\} \subset \mathbb{R}^n$ der Spaltenraum von $\boldsymbol{X}$,
$\Pi_L : \mathbb{R}^n \to L$ die orthogonale Projektion auf L
Für $\boldsymbol{y} \in \mathbb{R}^n$:

(i) $\Pi_L \boldsymbol{y} \in L$ und $\|\boldsymbol{y} - \Pi_L \boldsymbol{y}\|_2^2 \leq \min_{\boldsymbol{u} \in L} \|\boldsymbol{y} - \boldsymbol{u}\|_2^2$

(ii) $\Pi_L \boldsymbol{y} \in L$ und $\boldsymbol{y} - \Pi_L \boldsymbol{y} \perp L$

(und $\Pi_L \boldsymbol{y}$ ist durch (i) und durch (ii) charakterisiert)

$\Leftrightarrow$(ii)' $\Pi_L \boldsymbol{y} \in L$ und $\boldsymbol{u}^T(\boldsymbol{y} - \Pi_L \boldsymbol{y}) = 0$ für $\boldsymbol{u}$ aus einer Basis von L
$\iff \boldsymbol{X}^T(\boldsymbol{y} - \Pi_L \boldsymbol{y}) = \boldsymbol{0}$

Es ist $\Pi_L \boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ für $\boldsymbol{y} \in \mathbb{R}^n$:
$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \in L$ und $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{0}$

[Details ggfs. an der Tafel]

$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ der kleinste-Quadrate-Schätzer,

$$\widehat{\boldsymbol{y}} := \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

die angepassten Werte der y-Beobachtungen ("gefittete" Werte), $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ die „Hut-Matrix",

$$\boldsymbol{r} := \boldsymbol{y} - \widehat{\boldsymbol{y}} \left(= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \right)$$

die „Residuen"
(die Abweichungen der Beobachtungen vom Modell, sozusagen der „unerklärte Rest")

Bem.:
Die Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ ist $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Satz von Gauß-Markov

Im linearen Modell (mit unkorrelierten, zentrierten ε_i mit derselben Varianz σ^2) gilt

1. Der kleinste-Quadrate-Schätzer $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ ist erwartungstreu für $\boldsymbol{\beta}^*$, d.h. $\mathbb{E}_{\boldsymbol{\beta}^*}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right] = \boldsymbol{\beta}^*$ (für jede Wahl des „wahren“ $\boldsymbol{\beta}^*$).
2. $\tau : \mathbb{R}^p \to \mathbb{R}$ eine lineare Kenngröße von $\boldsymbol{\beta}$ (d.h. $\tau(\boldsymbol{\beta}) = \boldsymbol{c}^T\boldsymbol{\beta}$ für ein $\boldsymbol{c} \in \mathbb{R}^p$, z.B. $\tau(\boldsymbol{\beta}) = \beta_1$), dann ist $\widehat{\tau} := \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ erwartungstreuer Schätzer für $\tau(\boldsymbol{\beta}^*)$ und hat unter allen erwartungstreuen linearen Schätzern für $\tau(\boldsymbol{\beta}^*)$ die kleinste Varianz.

Weiterhin ist (Π_L=orth. Proj. auf den Spaltenraum von $\boldsymbol{X}$)

$$\widehat{\sigma}^2 := \frac{\|\boldsymbol{r}\|_2^2}{n-p} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\|_2^2}{n-p} = \frac{\|\boldsymbol{y} - \Pi_L\boldsymbol{y}\|_2^2}{n-p} = \frac{\|\boldsymbol{y}\|_2^2 - \|\Pi_L\boldsymbol{y}\|_2^2}{n-p}$$

ein erwartungstreuer Schätzer für die Varianz σ^2 (der ε_i).

[Details ggfs. an der Tafel]

Inhalt

Exkurs/Erinnerung: Rund um die multivariate Normalverteilung

$$\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

hat Dichte

$$\begin{aligned}\varphi_Z(z_1, \ldots, z_n) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-z_i^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2\sigma^2} \|z\|_2^2 \right)\end{aligned}$$

Die Verteilung von $\boldsymbol{Z}$ ist invariant unter orthogonalen Transformationen
(insbesondere: Projektionen auf verschiedene, zueinander orthogonale Teilräume sind unabhängig)

Rund um die multivariate Normalverteilung, 2

Seien $Z_1, \ldots, Z_m$, $Z'_1, \ldots, Z'_n$ u.a. Standard-normalverteilt.

- $Z_1^2 + \cdots + Z_m^2$ ist χ^2_m-verteilt

  (Dichte $\frac{(1/2)^{m/2}}{\Gamma(m/2)} x^{m/2-1} e^{-(1/2)x}$ auf $(0,\infty)$ $\quad$ $[\chi^2_m = \Gamma(m/2, 1/2)]$)

- $F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m Z_i^2}{\frac{1}{n}\sum_{j=1}^n Z_j'^2}$ ist Fisher$_{m,n}$-verteilt

  (Dichte $f_{m,n}(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{(m+n)}{2}}}$ auf $(0,\infty)$)

- $T := \frac{Z_1}{\sqrt{\frac{1}{n}\sum_{j=1}^n Z_j'^2}}$ ist Student-t_n-verteilt

  (Dichte $\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})} \frac{1}{\sqrt{n}} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$ auf $(-\infty,\infty)$)

ML-Schätzung im normalen Modell

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

Wir nehmen nun zusätzlich an, dass die ε_i u.i.v., $\sim \mathcal{N}(0,1)$ sind.

Gegeben $\boldsymbol{X}$ (und festes β) hat $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ dann die Dichte

$$\begin{aligned}
&\prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-(y_i - (\boldsymbol{X}\beta)_i)^2/(2\sigma^2)\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2/(2\sigma^2)\right)
\end{aligned}$$

Demnach ist $\widehat{\beta}_{\mathrm{LS}}$ auch der Maximum-Likelihood-Schätzer für β (und es ist $\widehat{\sigma}^2_{\mathrm{ML}} = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2/n$).

Verteilungseigenschaften (im normalen Modell)

1. $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$
2. $\frac{n-p}{\sigma^2}\widehat{\sigma}^2$ ist χ^2_{n-p}-verteilt und unabhängig von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$.
3. $\frac{1}{\sigma^2}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}^*)\|_2^2 = \frac{1}{\sigma^2}\|\Pi_L\boldsymbol{y} - \mathbb{E}_{\boldsymbol{\beta}^*}[\boldsymbol{y}]\|_2^2$ ist χ^2_p-verteilt und unabhängig von $\widehat{\sigma}^2$, insbesondere

$$\frac{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}^*)\|_2^2}{p\,\widehat{\sigma}^2} \sim \mathrm{Fisher}_{p,n-p}$$

4. $H \subset L$ linearer Teilraum von L, $\dim H = r < p$ und $\boldsymbol{X}\boldsymbol{\beta}^* \in H$, dann ist $\frac{1}{\sigma^2}\|\Pi_L\boldsymbol{y} - \Pi_H\boldsymbol{y}\|_2^2 \sim \chi^2_{p-r}$ und unabhängig von $\widehat{\sigma}^2$, insbesondere ist

$$\frac{n-p}{p-r}\frac{\|\Pi_L\boldsymbol{y} - \Pi_H\boldsymbol{y}\|_2^2}{\|\boldsymbol{y} - \Pi_L\boldsymbol{y}\|_2^2} = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \Pi_H\boldsymbol{y}\|_2^2}{(p-r)\widehat{\sigma}^2} \sim \mathrm{Fisher}_{p-r,n-p}.$$

Korollar: Konfidenzbereiche

Sei $\alpha \in (0,1)$.

1. $C_{\widehat{\beta}_{\mathrm{LS}}} := \{\tilde{\beta} \in \mathbb{R}^p : \|\boldsymbol{X}(\tilde{\beta} - \widehat{\beta}_{\mathrm{LS}})\|_2^2 < p\,\widehat{\sigma}^2\, f_{p,n-p;1-\alpha}\}$ ist ein Konfidenzbereich für β^* zum Sicherheitsniveau $1-\alpha$, wo $f_{p,n-p;1-\alpha}$ = $(1-\alpha)$-Quantil der Fisher$_{p,n-p}$-Verteilung.
2. $\tau(\beta) = \boldsymbol{c}^T\beta$ (mit einem $\boldsymbol{c} \in \mathbb{R}^p$) ein lineares Parametermerkmal.

$$C_{\tau(\widehat{\beta}_{\mathrm{LS}})} = \left(\boldsymbol{c}^T\widehat{\beta}_{\mathrm{LS}} - \delta\sqrt{\widehat{\sigma}^2}, \boldsymbol{c}^T\widehat{\beta}_{\mathrm{LS}} + \delta\sqrt{\widehat{\sigma}^2}\right),$$

mit $\delta = \sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}\, t_{n-p,1-\frac{\alpha}{2}}$ und $t_{n-p,1-\frac{\alpha}{2}}$ = $(1-\frac{\alpha}{2})$-Quantil der Student-T-Verteilung mit $n-p$-Freiheitsgraden, ist ein Konfidenzintervall für $\tau(\beta^*)$ zum Sicherheitsniveau $1-\alpha$.

Bem.: $C_{\widehat{\sigma}^2} := \left(\frac{n-p}{\chi^2_{n-p,1-\frac{\alpha}{2}}}\widehat{\sigma}^2, \frac{n-p}{\chi^2_{n-p,\frac{\alpha}{2}}}\widehat{\sigma}^2\right)$ ist ein Konfidenzintervall für σ^2 zum Sicherheitsniveau $1-\alpha$.

Inhalt

Prostata-Krebs-Daten von Stamey et al

Datensatz beschrieben in T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning*, Example 3.2.1, `http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`

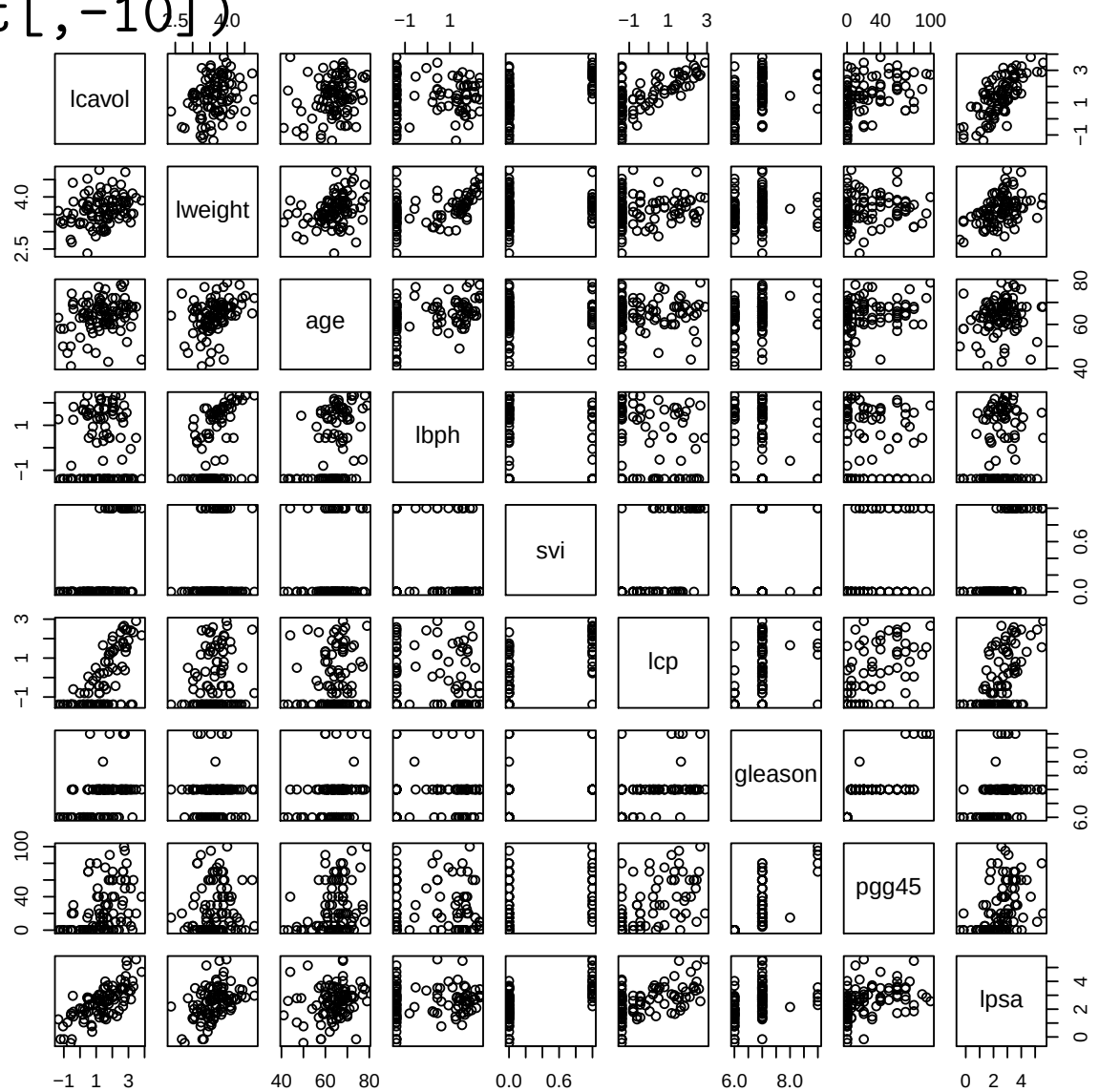
$n = 97$ Beobachtungen (Prostata-Krebs-Patienten),
$p = 8$ erklärende Variablen (plus "intercept")

> *log cancer volume* (`lcavol`), *log prostate weight* (`lweight`), `age`, *log of the amount of benign prostatic hyperplasia* (`lbph`), *seminal vesicle invasion* (`svi`), *log of capsular penetration* (`lcp`), *Gleason score* (`gleason`), *percent of Gleason scores 4 or 5* (`pgg45`)

Zielgröße ("response") *log prostate specific antigen* `lpsa`

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients, Journal of Urology 16: 1076–1083

```
> dat <- read.table("prostate.data", header=T)
> pairs(dat[,-10])
```



Zentrierung/Standardisierung

Wir folgen Hastie, Tibshirani & Friedman (und „üblicher Praxis“):

Zentrierung der Spalten von $\boldsymbol{X}$:
ersetze $x_{i,j}$ durch $x_{i,j} - \overline{x}_j$ mit $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$

Standardisierung der Spalten von $\boldsymbol{X}$: Skaliere so, dass $\frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 = 1$ für $j = 1, \ldots, p$
(d.h. ersetze $x_{i,j}$ durch $x_{i,j}/(\frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2)^{1/2}$, dann hat $\boldsymbol{X}^T\boldsymbol{X}$ auf der Diagonale 1en stehen)

```
> dat <- read.table("prostate.data", header=T)
> x <- dat[,1:8]
> xp <- scale(x,TRUE,TRUE)
> sdat <- data.frame(xp, lpsa=dat$lpsa)
```

Es gibt nicht-triviale Korrelationen zwischen den Spalten von $\boldsymbol{X}$:

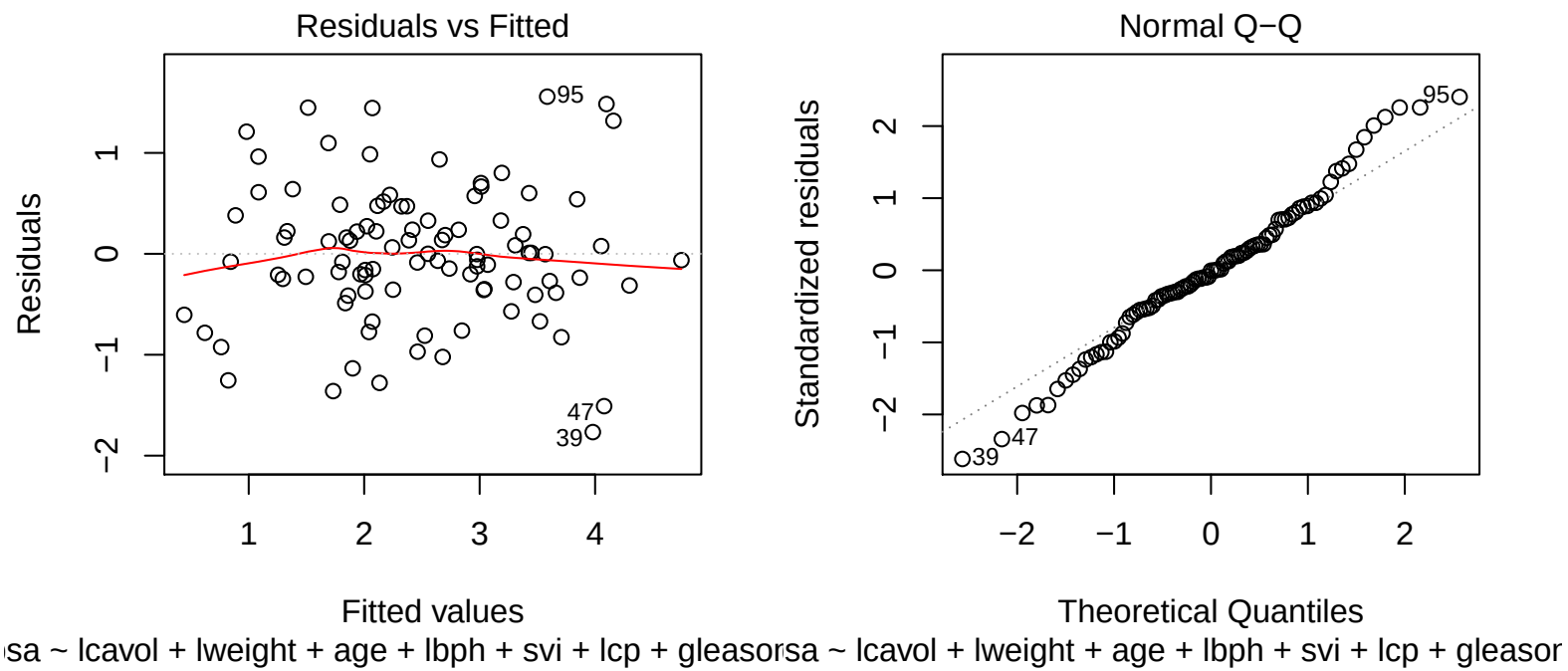
```
> round(cor(sdat[1:8]), digits=3)
```

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lcavol	1.000	0.281	0.225	0.027	0.539	0.675	0.432	0.434
lweight	0.281	1.000	0.348	0.442	0.155	0.165	0.057	0.107
age	0.225	0.348	1.000	0.350	0.118	0.128	0.269	0.276
lbph	0.027	0.442	0.350	1.000	-0.086	-0.007	0.078	0.078
svi	0.539	0.155	0.118	-0.086	1.000	0.673	0.320	0.458
lcp	0.675	0.165	0.128	-0.007	0.673	1.000	0.515	0.632
gleason	0.432	0.057	0.269	0.078	0.320	0.515	1.000	0.752
pgg45	0.434	0.107	0.276	0.078	0.458	0.632	0.752	1.000

```
> summary(modell)
Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45, data = sdat)
Residuals:
     Min       1Q   Median       3Q      Max
-1.76644 -0.35510 -0.00328  0.38087  1.55770
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47839    0.07102  34.895  < 2e-16 ***
lcavol       0.66515    0.10352   6.425 6.55e-09 ***
lweight      0.26648    0.08607   3.096  0.00263 **
age         -0.15820    0.08252  -1.917  0.05848 .
lbph         0.14031    0.08402   1.670  0.09848 .
svi          0.31533    0.09985   3.158  0.00218 **
lcp         -0.14829    0.12566  -1.180  0.24115
gleason      0.03555    0.11218   0.317  0.75207
pgg45        0.12572    0.12312   1.021  0.31000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Betrachten wir die Anpassung graphisch sowie einen Q-Q-Plot der Residuen

```
> plot(modell)
```



Wird der Fit schlechter, wenn wir `age`, `lcp`, `gleason` und `pgg45` weglassen (z.B. weil wir glauben, dass sie in Wirklichkeit keinen Einfluss haben, d.h. die entsprechenden $\beta_i =$ sind)?

```
> summary(modell2)
Call:
lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = sdat

Residuals:
     Min       1Q   Median       3Q      Max
-1.85483 -0.35990 -0.01492  0.44467  1.53051
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47839    0.07148  34.673  < 2e-16 ***
lcavol       0.62289    0.08781   7.094 2.63e-10 ***
lweight      0.22964    0.08415   2.729  0.00761 **
lbph         0.11403    0.08139   1.401  0.16454
svi          0.29207    0.08610   3.392  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Wird der Fit schlechter, wenn wir `age`, `lcp`, `gleason` und `pgg45` weglassen (z.B. weil wir glauben, dass sie in Wirklichkeit keinen Einfluss haben, d.h. die entsprechenden $\beta_i =$ sind)?

```
> anova(modell2, modell)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + lbph + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp
    pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     92 45.595
2     88 43.058  4     2.537 1.2963 0.2777
```

Bemerkung: ℓ_2-Regularisierung, Ridge regresssion

$\lambda \geq 0$ ein „Regularisierungsparameter,“

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) := \operatorname{argmin}\left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

Für $\lambda > 0$ ist $\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)$ (im Gegensatz zu $\widehat{\boldsymbol{\beta}}_{\text{LS}} = \widehat{\boldsymbol{\beta}}_{\text{ridge}}(0)$) nicht erwartungstreu,
kann aber ggfs. kleinere (quadratische) „Verlustfunktion“

$$\mathbb{E}_{\boldsymbol{\beta}^*}\left[\|\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}^*\|_2^2 \right]$$

haben (“bias-variance trade-off”), speziell wenn die Designmatrix $\boldsymbol{X}$ (einige) stark korrelierte Spalten hat.

Beachte: Man läßt typischerweise den “intercept” weg, da der Regularisierungsterm $\lambda\|\boldsymbol{\beta}\|_2^2$ die „Verschiebungsinvarianz“ des Schätzers bricht.

Die Berechnung von $\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)$ ist eine quadratische Minimierungsaufgabe, die wieder (im Prinzip) explizit gelöst werden kann:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

(mit $\boldsymbol{I} = p \times p$-Einheitsmarix)

Literaturhinweise

- Überblick z.B.: T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning*, Springer, 2009, speziell Ch. 3
- Für lineares Modell mit normalverteiltem Rauschen z.B.: H.O. Georgii, *Stochastik*, de Gruyter, 2002, G. Kersting, A. Wakolbinger, *Elementare Stochastik*, Birkhäuser, 2008
- Monographie: J.S. Stapleton, *Linear statistical models*, Wiley, 1995