

Stochastik für Informatik SoSe 2023

Lineare Modelle

Hanno Gottschalk

July 11, 2023

Regression mit mehreren Einflussgrößen	3
Endemische Pflanzen auf Galapagos	4
Galapagos - Fortsetzung	5
Galapagos: Visualisierung	6
Galapagos: Modellierung	7
Galapagos: Fit	8
Galapagos: Diagnostics	9
Die Modell-Matrix	10
Matrixschreibweise	11
Matrixschreibweise II	12
Matrixschreibweise III	13
Matrixschreibweise bei d Regressoren	14
Kleinste Quadrate Schätzer	15
Kleinste Quadrate Ansatz für lin. Modelle	16
Berechnung von $\hat{\beta}$	17
Kleinste Quadrate fit als Projektion	18
Projektion als Minimierung	19
Beispiele für Projektionen	20
Anwendung auf Lineare Modelle	21
Anw. auf Lin. Mod. – Beweis	22
Streuerlegung für lineare Modelle	23
Gesamte, Erklärte und Reststreuung	24
Streuerlegungssatz für lin. Modelle	25
Beweis Streuerlegung	26
Lineare Modelle mit nicht linearen Funktionen	27
Bremsweg revisited	28
Diagnostische Plots Bremsweg revisited	29
Definition: Lineares Modell	30
Lineare Modelle - Bemerkungen	31

Lineare Modell - Spezialfälle	32
---	----

Inhaltsverzeichnis der Vorlesung

- Regression mit zwei Einflussgrößen
- Die Modell-Matrix
- Kleinste Quadrate Schätzer
- Geometrische Interpretation des LSF
- Streuzerlegung
- Lineare Modelle mit nicht linearen Funktionen

- ## Inhaltsverzeichnis der Vorlesung
- Regression mit zwei Einflussgrößen
 - Die Modell-Matrix
 - Kleinste Quadrate Schätzer
 - Geometrische Interpretation des LSF
 - Streuzerlegung
 - Lineare Modelle mit nicht linearen Funktionen



Hanno Gottschalk Stochastik für Info – 2 / 32

Hanno Gottschalk Stochastik für Info – 2 / 32

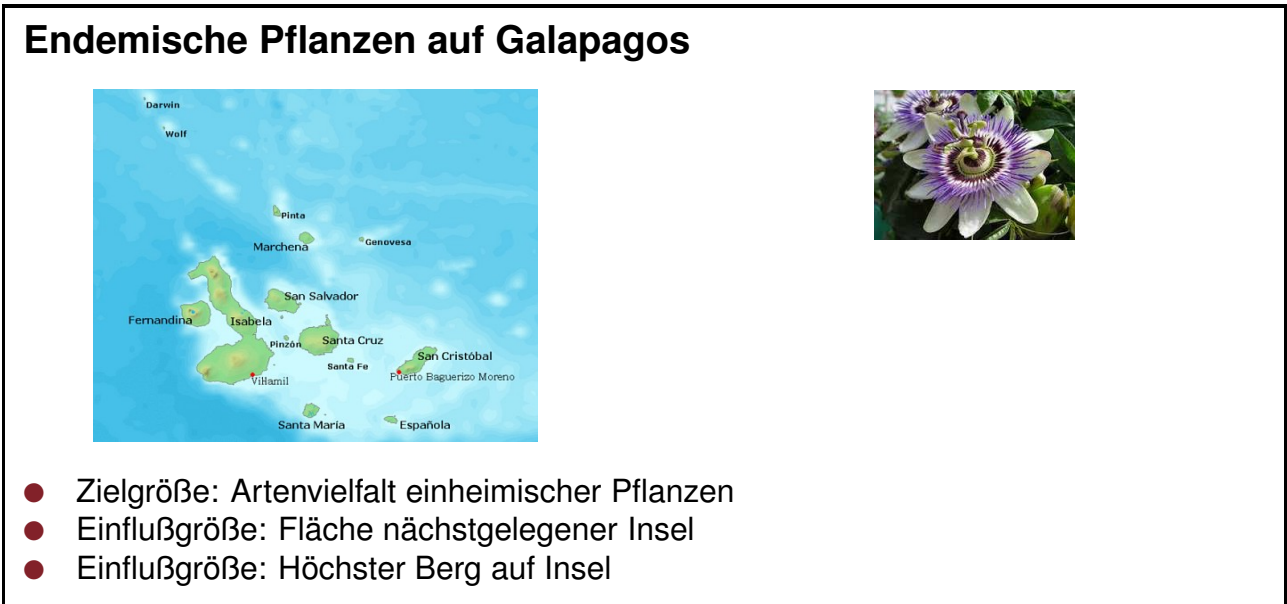
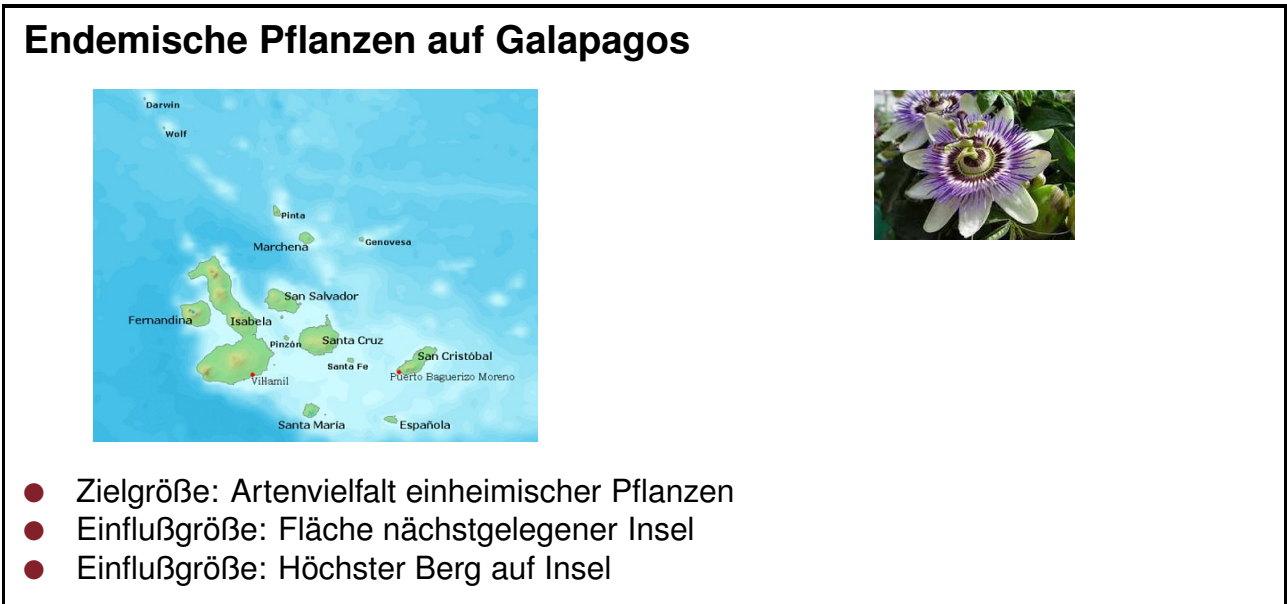
Regression mit mehreren Einflussgrößen 3 / 32



Regression mit mehreren Einflussgrößen 3 / 32

Endemische Pflanzen auf Galapagos



- Zielgröße: Artenvielfalt einheimischer Pflanzen
- Einflußgröße: Fläche nächstgelegener Insel
- Einflußgröße: Höchster Berg auf Insel



- # Endemische Pflanzen auf Galapagos
- 
- 
- Zielgröße: Artenvielfalt einheimischer Pflanzen
 - Einflußgröße: Fläche nächstgelegener Insel
 - Einflußgröße: Höchster Berg auf Insel

Hanno Gottschalk Stochastik für Info – 4 / 32

Hanno Gottschalk Stochastik für Info – 4 / 32

Galapagos - Fortsetzung

- i : Stat. Einheit – die Insel (30 Inseln)
- Y : Zielgröße – Anzahl beobachteter einheimischer Species
- X_1 Einflußgröße – Höchste Erhebung [m]
- X_2 Einflußgröße – Größe Nachbarinsel [km²]



	Species	Elevation	Adjacent
1			
2	Baltra	58	3481.84
3	Bartolome	31	109572.33
4	Caldwell	3	1140.78
5	Champion	25	480.18
6	Coamano	2	77903.82
7	Daphne Major	18	1191.84
8	Daphne Minor	24	930.34
9	Darwin	10	1682.85
10	Eden	8	7117.95
11	Enderby	2	1120.1
12	Espanola	97	1980.57
13	Fernandina	93	14944669.32
14	Gardner1	58	4958.27
15	Gardner2	5	2270.21
16	Genovesa	40	76129.49
17	Isabela	347	1707634.49
18	Marchena	51	34359.56
19	Onslow	2	250.1
20	Pinta	104	777129.49
21	Pinzon	108	4580.03
22	Las Plazas	12	9425.09
23	Rabida	70	367572.33
24	SanCristobal	280	7160.57
25	SanSalvador	237	9064.89

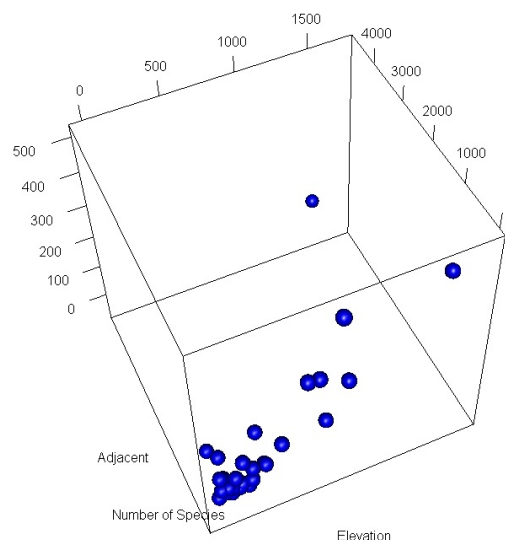
M. P. Johnson and P. H. Raven (1973) "Species number and endemism: The Galapagos Archipelago revisited" Science, 179, 893-895 - aus R-package faraway

Hanno Gottschalk

Stochastik für Info – 5 / 32

Galapagos: Visualisierung

3D Streuplot:



Hanno Gottschalk

Stochastik für Info – 6 / 32

Galapagos: Modellierung

Wir wählen einen linearen Ansatz für beide Einflußfaktoren:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \text{stat. Schwankungen} \quad (1)$$

Residuen:

$$\epsilon_i = \epsilon_i(\alpha, \beta_1, \beta_2) = y_i - (\alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2}) \quad (2)$$

Bestimme α, β_1, β_2 so dass

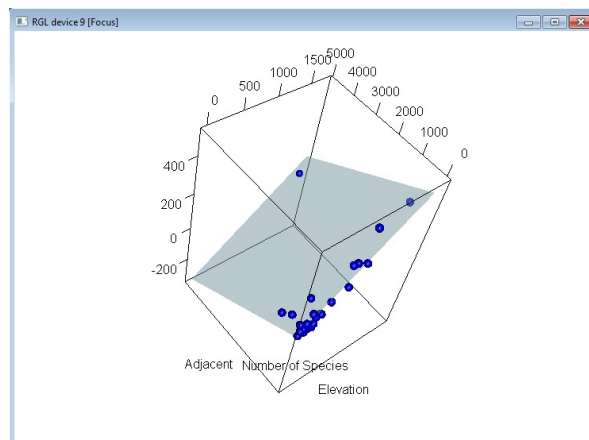
$$Q(\alpha, \beta_1, \beta_2) = \sum_{i=1}^n [\epsilon_i(\alpha, \beta_1, \beta_2)]^2 \longrightarrow \min \quad (3)$$

Die so gefundenen Parameter $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$ definieren eine Ebene im 3D Streuplot

Hanno Gottschalk

Stochastik für Info – 7 / 32

Galapagos: Fit

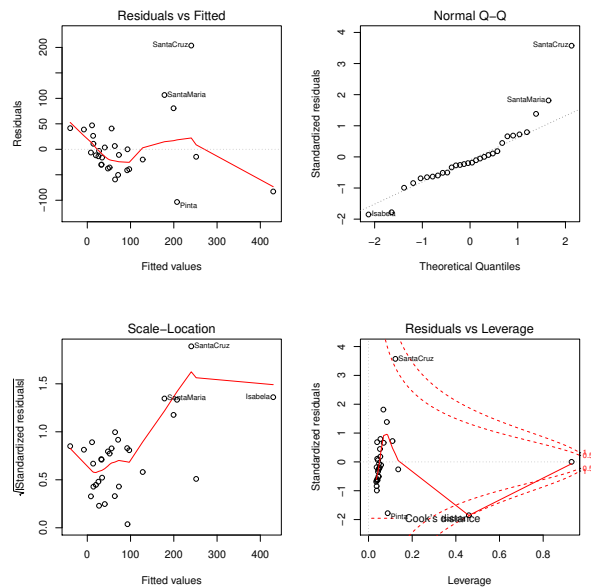


- $\hat{\alpha} = 1.4328$ (Intercept)
- $\hat{\beta}_1 = 0.2765/[m]$ (Elevation/Höhe)
- $\hat{\beta}_2 = -0.06889/[km]$ (Adjacent/Größe)

Hanno Gottschalk

Stochastik für Info – 8 / 32

Galapagos: Diagnostics



Hanno Gottschalk

Stochastik für Info – 9 / 32

Die Modell-Matrix

10 / 32

Matrixschreibweise

Wir schreiben alle Gleichungen auf einmal hin. . .

$$\begin{aligned} y_1 &= \alpha + x_{1,1}\beta_1 + x_{1,2}\beta_2 + \epsilon_1 \\ y_2 &= \alpha + x_{2,1}\beta_1 + x_{2,2}\beta_2 + \epsilon_2 \\ \dots &\vdots \quad \dots \\ y_n &= \alpha + x_{n,1}\beta_1 + x_{n,2}\beta_2 + \epsilon_n \end{aligned} \tag{4}$$

In Matrixschreibweise

$$\underline{y} = \alpha \underline{1} + \underline{\underline{M}} \underline{\beta} + \underline{\epsilon}, \quad \underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \tag{5}$$

$$\underline{\underline{M}} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}, \quad \underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \underline{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Hanno Gottschalk

Stochastik für Info – 11 / 32

Matrixschreibweise II

$$\underline{\underline{M}} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}$$

Die Matrix entspricht gerade den Regressorspalten in der Urliste



	Species	Elevation	Adjacent
1	Balta	58	3481.84
2	Barro Colorado	31	108572.33
3	Caldwell	3	1140.78
4	Champion	25	480.18
5	Coastal	2	77903.82
6	Daphne Major	18	1191.84
7	Daphne Minor	24	830.34
8	Darwin	10	1682.85
9	Eden	8	7117.85
10	Enderby	2	1120.1
11	Esperanza	97	1980.57
12	Fernandina	93	1494.669.32
13	Gardner1	58	4958.27
14	Gardner2	5	2270.21
15	Genovesa	40	76129.49
16	Isabela	347	1707634.49
17	Marthena	51	34359.56
18	Orinoco	2	250.1
19	Pinta	104	777129.49
20	Princeton	108	4580.03
21	Las Plazas	12	8425.09
22	Rabida	70	367572.33
23	San Cristobal	280	7160.57
24	San Salvador	237	9064.89

Hanno Gottschalk

Stochastik für Info – 12 / 32

Matrixschreibweise III

$$\underline{y} = \alpha \underline{1} + \underline{\underline{M}} \underline{\beta} + \underline{\epsilon}, \quad \underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \underline{\underline{M}} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}$$

\Leftrightarrow

$$\underline{y} = \underline{\underline{M}} \underline{\beta} + \underline{\epsilon}, \quad \underline{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \underline{\underline{M}} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix} \quad (6)$$

Hanno Gottschalk

Stochastik für Info – 13 / 32

Matrixschreibweise bei d Regressoren

Gegeben sei das statistische Modell

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_d X_d + \text{stat. Schwankungen} \quad (7)$$

wobei n Beobachtungen für Y vorliegen – y_1, \dots, y_n – und entsprechend n Beobachtungen für jeden Regressor X_j – $x_{1,j}, \dots, x_{n,j}$ – $j = 1, \dots, d$.

Dann lautet die Matrixschreibweise für dieses Modell

$$\underline{y} = \underline{\underline{M}} \underline{\beta} + \underline{\epsilon}, \quad \underline{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}, \quad \underline{\underline{M}} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \quad (8)$$

Def.: (8) definiert das Modell der multivariaten linearen Regression. Dies ist ein Spezialfall eines linearen Modells.

Hanno Gottschalk

Stochastik für Info – 14 / 32

Kleinste Quadrate Schätzer

15 / 32

Kleinste Quadrate Ansatz für lin. Modelle

Def. gegeben sei ein statistisches Modell

$$\underline{y} = \underline{\underline{M}} \underline{\beta} + \underline{\epsilon}$$

mit einer Modell-Matrix aus Observationen $\underline{\underline{M}}$ mit Rang $d + 1$ (falls $n \geq d + 1$: voller Rang) und der beobachteten Response \underline{y} .

Der kleinste Quadrate Schätzer $\hat{\underline{\beta}}$ ist eindeutig gegeben als Lösung des Problems

$$Q(\underline{\beta}) = \underline{\epsilon}' \underline{\epsilon} = (\underline{y} - \underline{\underline{M}} \underline{\beta})' (\underline{y} - \underline{\underline{M}} \underline{\beta}) \longrightarrow \min \quad (9)$$

Hanno Gottschalk

Stochastik für Info – 16 / 32

Berechnung von $\hat{\underline{\beta}}$

Wir suchen zunächst die Extrema von $Q(\underline{\beta})$:

$$\begin{aligned}\nabla Q(\underline{\beta}) &= -\underline{M}'(\underline{y} - \underline{M}\underline{\beta}) - \left[(\underline{y} - \underline{M}\underline{\beta})' \underline{M}\right]' \\ &= -[\underline{M}'\underline{y} - \underline{M}'\underline{M}\underline{\beta}] - [\underline{M}'\underline{y} - \underline{M}'\underline{M}\underline{\beta}] \\ &\stackrel{!}{=} 0\end{aligned}\tag{10}$$

\Leftrightarrow

$$\underline{M}'\underline{y} - \underline{M}'\underline{M}\underline{\beta} \stackrel{!}{=} 0\tag{11}$$

\Rightarrow Falls \underline{M} vollen Rang hat, liegt das einzige Extremum bei

$$\hat{\underline{\beta}} = (\underline{M}'\underline{M})^{-1} \underline{M}'\underline{y}\tag{12}$$

Es ist ein Minimum, da $\nabla^2 Q(\underline{\beta}) = \underline{M}'\underline{M}$ positiv definit ist.

Hanno Gottschalk

Stochastik für Info – 17 / 32

Kleinste Quadrate fit als Projektion

18 / 32

Projektion als Minimierung

Haben lineare Modelle mit mitteln der lin. Algebra aufgestellt. Begriffsbildungen in der lin. Algebra haben oft eine *geometrische Interpretation*. Nach dieser suchen wir hier für die lin. Modelle.

Gegeben sei ein Punkt $\underline{y} \in \mathbb{R}^n$ und ein q -dimensionaler linearer Unterraum $M \subset \mathbb{R}^n$.

Wiederholung: $M \subseteq \mathbb{R}^n$ linearer Unterraum $\Leftrightarrow \forall u, v \in M, a, b \in \mathbb{R}$ gilt $au + bv \in M$.

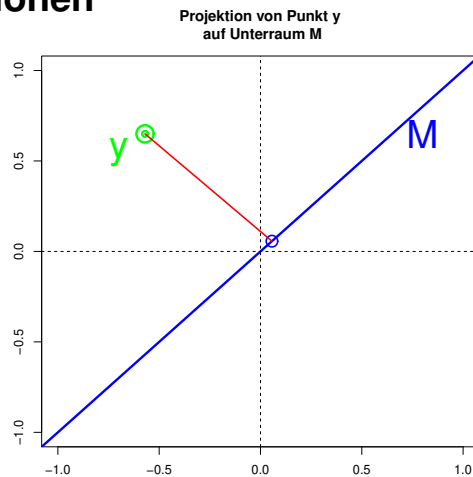
Es sei $\Pi_M \underline{y}$ die Projektion von \underline{y} auf M der Vektor in M mit dem kleinsten euklidischen Abstand zu \underline{y}

$$\Pi_M \underline{y} := \{\underline{u} \in M : |\underline{y} - \underline{u}|^2 = \sum_{i=1}^n (y_i - u_i)^2 \leq |\underline{y} - \underline{v}|^2 \forall \underline{v} \in M\}$$

Hanno Gottschalk

Stochastik für Info – 19 / 32

Beispiele für Projektionen



Pythagoras: $|\underline{y}|^2 = |\Pi_M \underline{y}|^2 + |\Pi_M \underline{y} - \underline{y}|^2$

Hanno Gottschalk

Stochastik für Info – 20 / 32

Anwendung auf Lineare Modelle

Def.: Gegeben sei die $n \times q$, $n \geq q$ Modell-Matrix $\underline{\underline{M}}$ eines linearen Modells mit maximalem Rang q . Dann ist der *Vorhersageraum* M des Modelles gegeben als Bildraum der mit $\underline{\underline{M}}$ assoziierten lin. Abbildung von \mathbb{R}^q nach \mathbb{R}^n .

$$M = \text{Bild}(\underline{\underline{M}}) = \{\underline{\underline{M}} \underline{\beta} : \underline{\beta} \in \mathbb{R}^q\} \quad (13)$$

Satz: Es sei \underline{y} der Vektor der beobachteten Zielgrößen eines lin. Modells $\underline{y} = \underline{\underline{M}} \underline{\beta} + \underline{\epsilon}$ mit Vorhersageraum $M \subseteq \mathbb{R}^n$. Dann gilt:

$$\Pi_M \underline{y} = \underline{\underline{M}} \hat{\underline{\beta}} \quad (14)$$

Hanno Gottschalk

Stochastik für Info – 21 / 32

Anw. auf Lin. Mod. – Beweis

Beweis:

Nach Def. minimiert $\hat{\underline{\beta}}$ die Residuenquadrate $|\underline{y} - \underline{M}\underline{\beta}|^2$.

Da $M = \{\underline{M}\underline{\beta} : \underline{\beta} \in \mathbb{R}^q\}$ minimiert $\underline{u} = \underline{M}\hat{\underline{\beta}}$ die quadrierte Euklidische Norm $|\underline{u} - \underline{y}|^2$ für $\underline{u} \in M$.

$\Rightarrow \underline{u} = \Pi_M \underline{y}$ per Definition von $\Pi_M \underline{y}$

qed.

Hanno Gottschalk

Stochastik für Info – 22 / 32

Streuerlegung für lineare Modelle

23 / 32

Gesamte, Erklärte und Reststreuung

Def: Gegeben sei das multivariate Regressionsmodell $\underline{y} = \underline{M}\underline{\beta} + \underline{\epsilon}$.

Die *Gesamtstreuung* SQT ist gegeben als

$$SQT = \sum_{j=1}^n (y_j - \bar{y})^2 \quad (15)$$

Die *Erklärte Streuung* SQE ist gegeben als

$$SQE = \sum_{j=1}^n (u_j - \bar{y})^2 \text{ mit } \underline{u} = \underline{M}\hat{\underline{\beta}} \quad (16)$$

Die *Reststreuung oder Residualstreuung* SQR ist gegeben als

$$SQR = \sum_{j=1}^n (u_j - y_j)^2 \text{ mit } \underline{u} = \underline{M}\hat{\underline{\beta}} \quad (17)$$

Hanno Gottschalk

Stochastik für Info – 24 / 32

Streuzerlegungssatz für lin. Modelle

Satz: Es sei $\underline{y} = \underline{M}\underline{\beta} + \underline{\epsilon}$ ein multivariates Regressionsmodell mit Intercept, also \underline{M} habe eine Spalte j mit allen Einträgen gleich 1. Dann gilt:

$$SQT = SQE + SQR \quad (18)$$

Bemerkung: Wie der folgende Beweis zeigt, genügt auch die Forderung

$$\underline{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in M = \text{Bild}(\underline{M}) \quad (19)$$

Hanno Gottschalk

Stochastik für Info – 25 / 32

Beweis Streuzerlegung

Beobachtung: Für $\underline{y} \in \mathbb{R}^n$ und $\underline{v} \in M$ gilt (LinA)

$$\Pi_M(\underline{y} - \underline{v}) = \Pi_M \underline{y} - \Pi_M \underline{v} = \Pi_M \underline{y} - \underline{v}$$

Nach Pythagoras:

$$SQT = |\underline{y} - \bar{y}\underline{1}|^2 = |\Pi_M(\underline{y} - \bar{y}\underline{1})|^2 + |\underline{y} - \bar{y}\underline{1} - \Pi_M(\underline{y} - \bar{y}\underline{1})|^2$$

Mit der Beobachtung und $\underline{1} \in M = \text{Bild}(\underline{M})$:

$$= |\Pi_M \underline{y} - \bar{y}\underline{1}|^2 + |\underline{y} - \bar{y}\underline{1} - \Pi_M \underline{y} + \bar{y}\underline{1}|^2 = |\Pi_M \underline{y} - \bar{y}\underline{1}|^2 + |\underline{y} - \Pi_M \underline{y}|^2$$

Mit dem Projektionssatz:

$$= |\underline{M}\hat{\underline{\beta}} - \bar{y}\underline{1}|^2 + |\underline{y} - \underline{M}\hat{\underline{\beta}}|^2 = SQE + SQR$$

qed.

Hanno Gottschalk

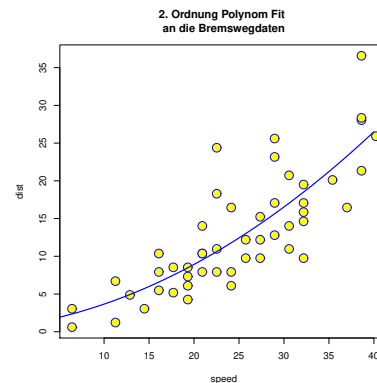
Stochastik für Info – 26 / 32

Bremsweg revisited

Ein angemesseneres Modell für die Abhängigkeit des Bremswegs von der Geschwindigkeit

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \text{stat. Schw.}$$

Hier übernimmt X^2 die Rolle von X_2

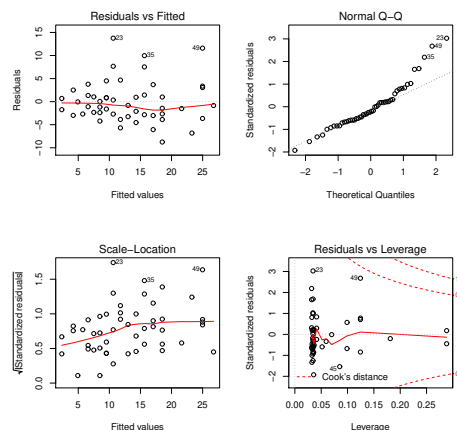


⇒ Lediglich die Modelmatrix $\underline{\underline{M}}$ muss anders aufgestellt werden, Lösungsweg für $\underline{\hat{\beta}}$ bleibt gleich

$$\underline{y} = \underline{\underline{M}}\underline{\beta} + \underline{\epsilon}, \quad \underline{\underline{M}} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad (20)$$

Hanno Gottschalk

Stochastik für Info – 28 / 32

Diagnostische Plots Bremsweg revisited

- Trends für quadratischen Fit ok ✓
- Streuung nimmt zu mit steigendem Bremsweg (nicht ganz befriedigend)

Hanno Gottschalk

Stochastik für Info – 29 / 32

Definition: Lineares Modell

Def.: In einer Stichprobe vom Umfang n seien für die Response Y die Werte y_1, \dots, y_n sowie die Regressoren $\underline{X} = (X_1, \dots, X_d)'$ die Werte $\underline{X}_1 = (x_{1,1}, \dots, x_{1,d}), \dots, \underline{X}_n = (x_{n,1}, \dots, x_{n,d})$ gemessen. Es seien $g_1, \dots, g_q : \mathbb{R}^d \rightarrow \mathbb{R}$ Funktionen.

Gegeben sei darüber hinaus der Ansatz

$$Y = \beta_1 g_1(\underline{X}) + \dots + \beta_q g_q(\underline{X}) + \text{stat. Schwankungen} \quad (21)$$

Dann kann man dieses lineare Modell in Matrixschreibweise aufstellen wie folgt:

$$\underline{y} = \underline{M} \underline{\beta} + \underline{\epsilon}, \quad \underline{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}, \quad \underline{M} = \begin{pmatrix} g_1(\underline{X}_1) & \dots & g_q(\underline{X}_1) \\ \vdots & \dots & \vdots \\ g_1(\underline{X}_n) & \dots & g_q(\underline{X}_n) \end{pmatrix} \quad (22)$$

Hanno Gottschalk

Stochastik für Info – 30 / 32

Lineare Modelle - Bemerkungen

Im linearen Modell ist NICHT die Abhängigkeit von den REGRESSOREN notwendigerweise linear, sondern die Abhängigkeit von den Koeffizienten!!!

$$Y = \alpha + \beta_1 X + \beta_2 \exp(X) + \text{stat. Schw.} \quad \text{IST lin. Modell!}$$

$$Y = \alpha + \beta_1 X + \exp(\beta_2 X) + \text{stat. Schw.} \quad \text{IST KEIN lin. Modell!}$$

Im engeren Sinne zählt noch die Normalverteilungsannahme für die stat. Schwankungen zum linearen Modell.

Hanno Gottschalk

Stochastik für Info – 31 / 32

Lineare Modell - Spezialfälle

Galapagos: Multivariate lin. Reg: $q = d + 1$, $g_1(\underline{X}) = 1$, $g_j(\underline{X}) = X_{j-1}$, $j = 1, \dots, d + 1$.

Bremsweg: Polynomiale Regression vom Grad $q - 1$ mit nur einem Merkmal: $d = 1$,
 $g_1(X) = 1, g_2(X) = X, \dots, q_q(X) = X^{q-1}$

Die Modellmatrix ändert sich, der Löser bleibt immer derselbe!!!

$$\underline{\hat{\beta}} = (\underline{\underline{M}}' \underline{\underline{M}})^{-1} \underline{\underline{M}}' \underline{y}$$