



Ruprecht-Karls-Universität Heidelberg

Institut für Angewandte Mathematik

Prof. Dr. Jan JOHANNES

STATISTIK 1

*Gliederung zur Vorlesung
im Wintersemester 2015/16*

vorläufige Fassung stand 14. Oktober 2015

Falls Sie **Fehler in der Gliederung** finden, teilen Sie mir diese bitte
per eMail an johannes@math.uni-heidelberg.de mit.

Im Neuenheimer Feld 294, 69120 Heidelberg

Telefon: +49 6221 54.62.76 – Fax: +49 6221 54.53.31

eMail: johannes@math.uni-heidelberg.de

Webseite zur Vorlesung: www.razbaer.eu/jan.johannes/vl/ST1-WS15/

Inhaltsverzeichnis

1	Statistische Inferenz im linearen Modell	1
1.1	Das lineare Modell	1
1.2	Methode der kleinsten Quadrate	5
1.3	Der Satz von Gauß-Markov	6
1.4	Die multivariate Normalverteilung	7
1.5	Das normale lineare Modell	12
1.6	Asymptotische Theorie	14
1.7	Residuenanalyse	15
2	Entscheidungstheorie	17
2.1	Formalisierung eines statistischen Problem	17
2.2	Minimax- und Bayes-Ansatz	20
2.3	Das Stein-Phänomen	24
3	Schätztheorie	27
3.1	Dominierte Modelle	27
3.2	Erschöpfende Statistik	27
3.3	Exponentialfamilien	31
3.4	Vollständige Statistik	33
3.5	Erwartungstreue Schätzer	34
3.6	Informationsungleichungen	36
3.7	Translationsäquivalente Schätzer	39
4	Allgemeine Schätzmethoden	41
4.1	Momentenschätzer	41
4.2	Maximum-Likelihood-Schätzer	43
4.3	Minimum-Kontrast-Schätzer	45
5	Testtheorie	51
5.1	Neyman-Pearson-Theorie	51
5.2	Bedingte Tests	54
5.3	Likelihood-Quotienten-Test	56

Kapitel 1

Statistische Inferenz im linearen Modell

1.1 Das lineare Modell

§1.1.1 **Beispiel.** In der folgenden Tabelle ist ein Auszug des „Cars93“ Datensatzes aus dem Statistikpaket R Core Team [2015] (library {MASS}) angegeben. Der Datensatz umfasst unter anderem den Preis, die Anzahl der Zylinder (Zyl.), den Hubraum (Hub.), die Breite sowie das Herkunftsland für 93 in den USA im Jahr 1993 verkauften Autos.

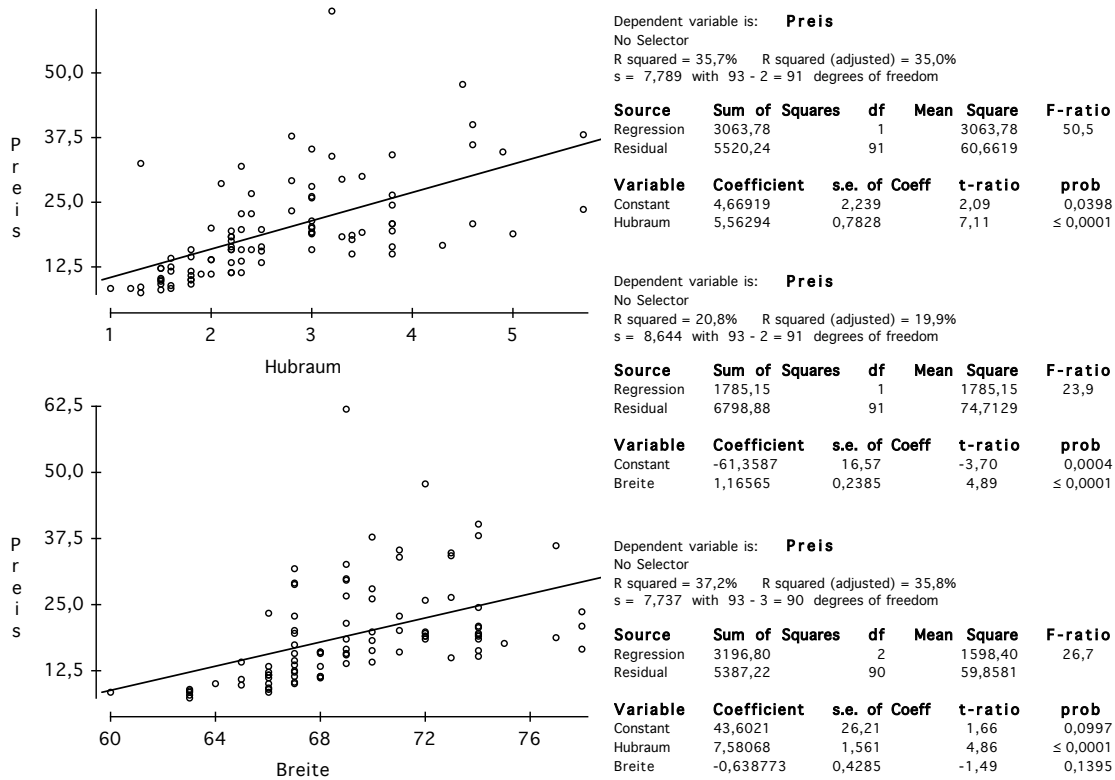
Preis	Zyl.	Hub.	Breite	Herkunft	Preis	Zyl.	Hub.	Breite	Herkunft
15.9	4	1.8	68	non-USA	10	4	1.5	64	non-USA
33.9	6	3.2	71	non-USA	13.9	4	2	69	non-USA
29.1	6	2.8	67	non-USA	47.9	8	4.5	72	non-USA
37.7	6	2.8	70	non-USA	28	6	3	70	non-USA
30	4	3.5	69	non-USA	35.2	6	3	71	non-USA
15.7	4	2.2	69	USA	34.3	6	3.8	73	USA
20.8	6	3.8	74	USA	36.1	8	4.6	77	USA
23.7	6	5.7	78	USA	8.3	4	1.6	66	non-USA
26.3	6	3.8	73	USA	11.6	4	1.8	66	non-USA
34.7	8	4.9	73	USA	16.5	4	2.5	69	non-USA
40.1	8	4.6	74	USA	19.1	6	3	72	non-USA
11.4	4	2.2	68	USA	31.9	4	2.3	67	non-USA
15.1	6	3.4	74	USA	61.9	6	3.2	69	non-USA
15.9	4	2.2	71	USA	14.1	4	1.6	65	USA
16.3	6	3.8	74	USA	14.9	6	3.8	73	USA
16.6	6	4.3	78	USA	10.3	4	1.5	67	non-USA

Preis, Anzahl der Zylinder (Zyl.), Hubraum (Hub.), Breite sowie Herkunftsland von in den USA verkauften Autos.

Sei Y_i der Preis des i -ten Autos mit Hubraum z_{1i} und Breite z_{2i} . Wir nehmen an, die Autos seien austauschbar und es existiert ein linearer Zusammenhang (vgl. nachfolgende Graphik) zwischen dem erwarteten Verhalten des Preises und den erklärenden Variablen Hubraum und Breite:

$$\mathbb{E}Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i}, \quad i = 1, \dots, 93.$$

Wir möchten statistische Aussagen über die Parameter β_1 und β_2 treffen, wie zum Beispiel die Werte der Parameter schätzen, Hypothesen der Form $\beta_1 = 0$ oder $\beta_2 = 0$ verifizieren oder den zu Grunde gelegten linearen Zusammenhang überprüfen.



Preis in Abhängigkeit des Hubraumes bzw. der Breite des Autos. □

§1.1.2 **Einfache lineare Regression.** Zu einem vorgegeben (nicht zufälligem) Versuchsplan $z_1, \dots, z_n \in \mathbb{R}$ beobachten wir Realisierungen der reellwertigen Zufallsvariablen (ZV'en)

$$Y_i = a + bz_i + \varepsilon_i, \quad i = 1, \dots, n,$$

wobei die zentrierten ZV'en $\{\varepsilon_i\}_{i=1}^n$ (d.h. $\mathbb{E}(\varepsilon_i) = 0$) Messfehler modellieren und $a, b \in \mathbb{R}$ unbekannte Parameter sind. Man denke z.B. an Messungen der Leitfähigkeit Y_i eines Stoffes in Abhängigkeit der Temperatur z_i , eines Effektes Y_i in Abhängigkeit einer Dosierung z_i oder eines Klausurergebnisses Y_i in Abhängigkeit der Klassengröße z_i . Offensichtlich gilt,

$$\mathbb{E}(Y_i) = a + bz_i, \quad i = 1, \dots, n.$$

so dass ein linearer Zusammenhang nur zwischen der erklärenden Variable x_i und der Erwartung der zu erklärenden zufälligen Größe Y_i zu Grunde gelegt wird. Betrachten wir weiterhin die n -dimensionalen zufälligen Vektoren $Y = (Y_1, \dots, Y_n)^t$ und $\varepsilon := (\varepsilon_1, \dots, \varepsilon_n)^t$, den unbekannt Parametervektor $\beta = (a, b)^t \in \mathbb{R}^2$ sowie die vorgegebene (Design-)Matrix $X = (x_1, \dots, x_n)^t \in \mathbb{R}^{n \times 2}$ mit Zeilen $x_i^t = (1, z_i)$, $i = 1, \dots, n$, dann lässt sich die einfache lineare Regression kompakt in der Form $Y = X\beta + \varepsilon$ schreiben. Wir bezeichnen weiterhin mit $\Sigma = \text{Cov}(\varepsilon) \in \mathbb{R}^{n \times n}$ die Kovarianzmatrix von ε , d.h. für den Eintrag Σ_{ij} in der i -ten Zeile und j -ten Spalte von $\Sigma := (\Sigma_{ij})_{1 \leq i, j \leq n}$ gilt $\Sigma_{ij} = \text{Cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j)$. □

§1.1.3 **Bemerkung.** Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, strikt positiv-definite Matrix ist. Insbesondere, ist dann Σ diagonalisierbar mit $\Sigma = U\Lambda U^t$ für eine Diagonalmatrix

$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ und eine unitäre Matrix U . Für $s \in \mathbb{R}$ setzen wir $\Sigma^s = U\Lambda^s U^t$ mit $\Lambda^s = \text{diag}(\lambda_1^s, \dots, \lambda_n^s)$. Wie erwartet, gilt $(\Sigma^{-1/2})^2 = \Sigma^{-1}$ und somit $\|\Sigma^{-1/2}v\|^2 = \langle \Sigma^{-1}v, v \rangle$. \square

§1.1.4 **Definition.** Ein **lineares Modell** beschreibt *adäquat* den Zusammenhang zwischen einem zu erklärenden, zufälligem Vektor (Zielgröße) $Y \in \mathbb{R}^n$ mit $\mathbb{E}\|Y\|^2 < \infty$ und einer erklärenden, vorgegebenen Matrix $X \in \mathbb{R}^{n \times p}$, der *Designmatrix* oder Matrix der Effekte, falls ein Parametervektor $\beta \in \mathbb{R}^p$ existiert, so dass $\mathbb{E}(Y) = X\beta$ gilt. Die Kovarianzmatrix $\Sigma = \text{Cov}(\varepsilon) \in \mathbb{R}^{n \times n}$ des zentrierten zufälligen Vektors $\varepsilon := Y - X\beta$, den *Fehler- oder Störgrößen*, sowie der Vektor $\beta \in \mathbb{R}^p$ sind unbekannte Parameter in einem linearem Modell. Beobachtet wird eine Realisierung von Y und die Designmatrix X und wir schreiben abkürzend $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$. In einem **gewöhnlichen linearen Modell** gilt weiterhin $\Sigma = \sigma^2 \text{Id}_n$ für ein Fehlerniveau $\sigma > 0$, wobei $\text{Id}_n \in \mathbb{R}^{n \times n}$ die Einheitsmatrix bezeichnet. \square

§1.1.5 **Beispiele.** (a) Ein zufälliger Vektor $Y \in \mathbb{R}^n$ folgt einem *Lokations-Skalen-Modell*, falls $\mathbb{E}(Y) = \mu \mathbb{1}_n$ mit $\mathbb{1}_n := (1, \dots, 1)^t \in \mathbb{R}^n$ und $\text{Cov}(Y) = \sigma^2 \text{Id}_n$ gilt. Die unbekannt Parameter sind $\mu \in \mathbb{R}$ als auch $\sigma > 0$. Wir schreiben abkürzend $Y \odot \{\mathcal{L}(\mu \mathbb{1}_n, \sigma^2 \text{Id}_n), \mu \in \mathbb{R}, \sigma > 0\}$. Sind die Koordinaten von Y zusätzlich unabhängige und identisch verteilte (u.i.v.) reellwertige ZV'en, so ist die Verteilung von Y durch das Produkt der eindimensionalen Randverteilungen gegeben und wir schreiben $Y \odot \{\mathcal{L}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$. Wird die Varianz σ_o^2 der Beobachtungen als bekannt vorausgesetzt, so erfüllt der zufällige Vektor Y ein *Lokations-Modell* und wir schreiben abkürzend $Y \odot \{\mathcal{L}(\mu \mathbb{1}_n, \sigma_o^2 \text{Id}_n), \mu \in \mathbb{R}\}$ oder $Y \odot \{\mathcal{L}^{\otimes n}(\mu, \sigma_o^2), \mu \in \mathbb{R}\}$. Wird dagegen der Erwartungswert μ_o als bekannt vorausgesetzt, so folgt der zufällige Vektor Y einem *Skalen-Modell* und wir schreiben abkürzend $Y \odot \{\mathcal{L}(\mu_o \mathbb{1}_n, \sigma^2 \text{Id}_n), \sigma > 0\}$ oder $Y \odot \{\mathcal{L}^{\otimes n}(\mu_o, \sigma^2), \sigma > 0\}$. Setzen wir $\beta = \mu$ und $X = \mathbb{1}_n$ so sind die drei Modelle offensichtlich (gewöhnliche) lineare Modelle.

(b) *Varianzanalyse mit einem Faktor.* Es werden q Proben an p Labore geschickt, wir erhalten zu jeder Probe einen Messwert, die wir als Realisierung von ZV'en

$$Y_{jk} = \mu_j + \varepsilon_{jk}, \quad j = 1, \dots, p, \quad k = 1, \dots, q,$$

auffassen. Ein Anordnen der ZV'en als $n = pq$ dimensionalen Vektor, $Y = (Y_1, \dots, Y_n)^t$ mit $Y_i = Y_{jk}$ und $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ mit $\varepsilon_i = \varepsilon_{jk}$ für $i = k + (j - 1)q$ erlaubt es uns, kompakt $Y = X\beta + \varepsilon$ zu schreiben, wobei $\beta := (\mu_1, \dots, \mu_p)^t$ und $X = \text{Id}_p \otimes \mathbb{1}_q$. Hier bezeichnet \otimes das Kronecker-Produkt, d.h. $A \otimes B := (a_{ij}B)$ für zwei Matrizen A und B . Insbesondere, folgt also der zufällige Vektor Y einem linearen Modell.

(c) Der Zusammenhang zwischen vorgegebenen Designpunkten $z_1, \dots, z_n \in \mathbb{R}$ und einem zufälligem Vektor $Y \in \mathbb{R}^n$ wird durch eine *polynomiale Regression* beschrieben, falls Parameter $a_0, \dots, a_{p-1} \in \mathbb{R}$ existieren, so dass

$$\mathbb{E}(Y_i) = a_0 + a_1 z_i + a_2 z_i^2 + \dots + a_{p-1} z_i^{p-1}, \quad i = 1, \dots, n,$$

gilt. Bezeichnen wir mit $\beta = (a_0, \dots, a_{p-1})^t$ den Vektor der unbekannt Parameter und mit

$$X = \begin{pmatrix} 1 & z_1 & z_1^2 & \dots & z_1^{p-1} \\ 1 & z_2 & z_2^2 & \dots & z_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \dots & z_n^{p-1} \end{pmatrix}$$

die Designmatrix vom Vandermonde-Typ, so gilt $\mathbb{E}(Y) = X\beta$ und es liegt somit ein lineares Modell vor. Für $p > 2$ ist der Zusammenhang zwischen den Designpunkten $\{z_i\}$ und den Beobachtungen $\{Y_i\}$ insbesondere nichtlinear. Auf Grund der linearen Abhängigkeit vom Parametervektor β wird das Modell linear genannt. Eine natürliche Verallgemeinerung der Modellierung eines nichtlinearer Zusammenhang zwischen den Designpunkten $\{z_i\}$ und den Beobachtungen $\{Y_i\}$ ist

$$\mathbb{E}(Y_i) = \beta_1\psi_1(z_i) + \cdots + \beta_p\psi_p(z_i), \quad i = 1, \dots, n,$$

mit unbekanntem Parametervektor $\beta = (\beta_1, \dots, \beta_p)^t$ und vorgegebene Basisfunktionen $\{\psi_j\}$, zum Beispiel Splinefunktionen. Setzen wir $X := (\psi_k(z_j))_{jk}$ so gilt erneut $\mathbb{E}(Y) = X\beta$ und das zugrunde liegende Modell ist linear. \square

§1.1.6 Definition. In einem linearen Modell $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ heißt der Parameter $\beta \in \mathbb{R}^p$ oder allgemeiner der abgeleitete Parameter $\gamma(\beta) \in \mathbb{R}^q$ für eine vorgegebene Funktion $\gamma : \mathbb{R}^p \rightarrow \mathbb{R}^q$ *identifizierbar*, falls $\mathbb{E}_{\beta_0} Y = \mathbb{E}_\beta Y$ impliziert $\gamma(\beta_0) = \gamma(\beta)$. \square

§1.1.7 Lemma. Sei $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ und $C \in \mathbb{R}^{q \times p}$ eine vorgegebene Matrix. Der *abgeleitete lineare Parameter* $\gamma(\beta) := C\beta \in \mathbb{R}^q$ ist genau dann identifizierbar wenn eine Matrix $A \in \mathbb{R}^{q \times n}$ existiert, so dass $C = AX$ gilt. \square

Beweis von Lemma §1.1.7. in der Vorlesung. \square

§1.1.8 Korollar. In einem linearen Modell $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ ist der Parameter $\beta \in \mathbb{R}^p$ genau dann identifizierbar, wenn die Designmatrix X den Rang $\text{rg}[X] = p$ besitzt. \square

Beweis von Korollar §1.1.8. in der Vorlesung. \square

§1.1.9 Bemerkung. Besitzt in einem linearen Modell die Designmatrix X den Rang $\text{rg}[X] = r < p$, so lässt sich durch eine geeignete Transformation $\gamma = C\beta$ und $\tilde{X} = XU$ für $C \in \mathbb{R}^{r \times p}$ und $U \in \mathbb{R}^{p \times r}$ erreichen, dass γ in dem reparametrisierten linearen Modell $\mathbb{E}Y = \tilde{X}\gamma$ identifizierbar ist. Dies ist genau dann der Fall, wenn $XUC = X$ und $\text{rg}[XU] = r$ gilt. \square

§1.1.10 Beispiele. (a) (*Einfache lineare Regression §1.1.2 fortgesetzt.*) Die Parameter a und b sind identifizierbar, falls mindestens zwei Effekte des Versuchsplans $\{z_i\}$ verschieden sind.

(b) (*Polynomiale Regression §1.1.5(c) fortgesetzt.*) Die Determinante einer Matrix vom Vandermonde-Typ ist im Fall $p = n$ gegeben durch $\prod_{p \geq k > j \geq 1} (x_k - x_j)$. Damit ist eine hinreichende und notwendige Bedingung für die Identifizierbarkeit des Parameters β , dass mindestens p verschiedene Effekte existieren. \square

§1.1.11 Bemerkung. Es gibt wichtige *Verallgemeinerungen linearer Modelle* (GLM für *Generalized Linear Model*). Der Zusammenhang zwischen einem zufälligem Vektor $Y \in \mathbb{R}^n$ und einer Designmatrix $X = (x_1, \dots, x_n)^t \in \mathbb{R}^{n \times p}$ ist durch ein *verallgemeinertes lineares Modell* mit vorgegebener Linkfunktion ℓ beschrieben, falls ein Parametervektor $\beta \in \mathbb{R}^p$ existiert, so dass $\mathbb{E}(Y_i) = \ell(x_i^t \beta)$, $i = 1, \dots, n$, gilt. Nehmen wir an, dass die ZV Y_i das Auftreten eines positiven oder negativen Effektes nach Verabreichung eines Medikamentes wiedergibt. In diesem Fall ist $Y_i \sim \mathfrak{B}\text{in}(1, \pi_i)$ eine Bernoulli-ZV und die Erfolgswahrscheinlichkeit π_i der unbekanntem Parameter. Eine *logistische Regression* liegt nun vor, falls ein Parametervektor

$\beta \in \mathbb{R}^p$ existiert, so dass $\log(\pi_i/(1 - \pi_i)) = x_i^t \beta$ oder äquivalent $\pi_i = \{1 + \exp(-x_i^t \beta)\}^{-1}$ für $i = 1, \dots, n$ gilt. Die Linkfunktion $\ell(x) = \{1 + \exp(-x)\}^{-1}$, $x \in \mathbb{R}$, entspricht gerade der logistischen Verteilungsfunktion, so dass wir auch von einem Logitmodell sprechen. Ein weiteres Beispiel, ist das Probitmodell, in dem ℓ der Verteilungsfunktion einer Standardnormalverteilung entspricht. \square

1.2 Methode der kleinsten Quadrate

Zur Erinnerung, im Sinne des mittleren quadratischen Fehlers (MSE für *mean squared error*) die beste konstante Approximation einer reellwertigen ZV Z mit $\mathbb{E}(Z^2) < \infty$ ist ihr Erwartungswert $\mu = \mathbb{E}(Z)$, d.h. $\mathbb{E}(Z - \mu)^2 = \min_{a \in \mathbb{R}} \mathbb{E}(Z - a)^2$. Das folgende Lemma verallgemeinert diesen Sachverhalt und motiviert zudem die Methode der kleinsten Quadrate.

§1.2.1 Lemma. In einem linearen Modell $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ gilt

$$\beta \in \arg \min_{b \in \mathbb{R}^p} \mathbb{E} \|\Sigma^{-1/2}(Y - Xb)\|^2$$

$$:\Leftrightarrow \mathbb{E} \|\Sigma^{-1/2}(Y - X\beta)\|^2 = \min_{b \in \mathbb{R}^p} \mathbb{E} \|\Sigma^{-1/2}(Y - Xb)\|^2. \quad (1.1)$$

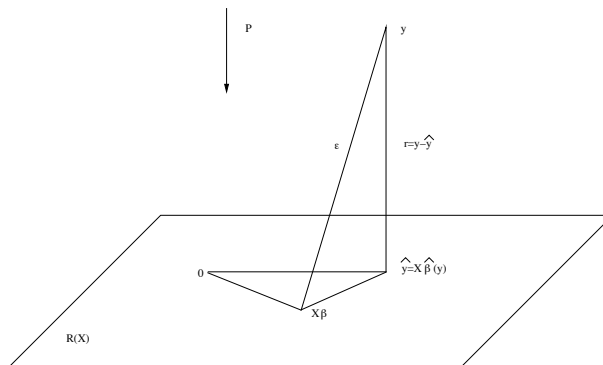
Beweis von Lemma §1.2.1. in der Vorlesung. \square

§1.2.2 Definition. In einem linearen Modell $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ heißt jede (messbare) Wahl von $\hat{\beta}$, so dass

$$\hat{\beta} \in \arg \min_{b \in \mathbb{R}^p} \|\Sigma^{-1/2}(Y - Xb)\|^2 \quad (1.2)$$

verallgemeinerter Kleinste-Quadrate-Schätzer (vKQS oder GLSE für *generalized least squares estimator*) des unbekanntem Parametervektors β . Im gewöhnlichen Fall ($\Sigma = \sigma^2 \text{Id}_n$) bezeichnen wir $\hat{\beta}$ als **gewöhnlichen Kleinste-Quadrate-Schätzer** (gKQS oder OLSE für *ordinary least squares estimator*). \square

§1.2.3 Geometrische Interpretation. Betrachten wir eine Realisierung y der Beobachtung Y als einen Punkt im n -dimensionalen Raum \mathbb{R}^n und variieren wir den Parameter β , so beschreibt $X\beta$ den k -dimensionalen Unterraum $\mathcal{R}(X)$, d.h. eine k -dimensionale Hyperebene durch den Ursprung im \mathbb{R}^n . Der gewöhnliche Kleinste-Quadrate-Schätzwert $\hat{\beta}(y)$ gibt uns nun den Punkt $X\hat{\beta}(y)$ auf der Hyperebene, der der Beobachtung y am nächsten liegt. Da die \mathcal{L}^2 -Norm durch ein Skalarprodukt $\langle \cdot, \cdot \rangle$ induziert ist, bedeutet die Wahl der \mathcal{L}^2 -Norm als Abstand im \mathbb{R}^n , geometrisch, dass wir y orthogonal bzgl. des Skalarproduktes $\langle \cdot, \cdot \rangle$ auf diese Hyperebene projizieren.



\square

§1.2.4 **Lemma.** Setze $\tilde{X} := \Sigma^{-1/2}X$ sowie $\tilde{Y} := \Sigma^{-1/2}Y$. Bezeichne mit $\mathcal{R}(\tilde{X}) := \{\tilde{X}b : b \in \mathbb{R}^p\}$ den Bildraum der linearen Abbildung \tilde{X} und mit $\Pi_{\mathcal{R}(\tilde{X})}$ die orthogonale Projektion von \mathbb{R}^n auf $\mathcal{R}(\tilde{X})$. Dann sind in einem linearen Modell $Y \odot \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ die folgenden Aussagen äquivalent: (i) $\hat{\beta}$ ist vKQS, d.h. $\hat{\beta}$ erfüllt (1.2), (ii) $\tilde{X}\hat{\beta} = \Pi_{\mathcal{R}(\tilde{X})}\tilde{Y}$, (iii) $\tilde{X}^t\tilde{X}\hat{\beta} = \tilde{X}^t\tilde{Y}$ („Normalgleichungen“). Insbesondere existiert der vKQS. \square

Beweis von Lemma §1.2.4. in der Vorlesung. \square

§1.2.5 **Korollar.** Sei X eine Designmatrix mit $\text{rg}[X] = p$, dann gilt $\Pi_{\mathcal{R}(\tilde{X})} = \tilde{X}(\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t$ und $\hat{\beta} = (\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t\tilde{Y} = (X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1}Y$ ist der eindeutige vKQS. Weiterhin ist im gewöhnlichen linearen Modell der gKQS $\hat{\beta} = (X^tX)^{-1}X^tY$ eindeutig und unabhängig von der Kenntnis von σ^2 . \square

§1.2.6 **Bemerkung.** Die Matrix $\tilde{X}^+ := (\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t$ heißt auch **Moore-Penrose-Inverse** von \tilde{X} und für die vKQS gilt $\hat{\beta} = \tilde{X}^+\tilde{Y}$. \square

§1.2.7 **Einfache lineare Regression** (§1.1.2 fortgesetzt). Wir wählen eine alternative Parametrisierung $\beta_1 := a + b\bar{z}$ sowie $\beta_2 := b$ mit $\bar{z} = n^{-1}\sum_{i=1}^n z_i$. Dann gilt

$$Y_i = \beta_1 + \beta_2(z_i - \bar{z}) + \varepsilon_i, \quad i = 1, \dots, n.$$

Setze weiterhin $x_i = (1, z_i - \bar{z})^t$, $i = 1, \dots, n$ und $X = (x_1, \dots, x_n)^t$, so dass $\mathbb{E}(Y) = X\beta$ mit $\beta = (\beta_1, \beta_2)^t$. Wir bestimmen im Folgenden einen gKQS von β , dazu setze $\bar{Y} := n^{-1}\sum_{i=1}^n Y_i$, $S_{zY} := \sum_{i=1}^n (z_i - \bar{z})Y_i = \sum_{i=1}^n (z_i - \bar{z})(Y_i - \bar{Y})$ und $S_{zz} := \sum_{i=1}^n (z_i - \bar{z})^2$, dann gilt

$$\begin{aligned} X^tY &= \sum_{i=1}^n x_i Y_i = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n (z_i - \bar{z})Y_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ S_{zY} \end{pmatrix} \\ X^tX &= \sum_{i=1}^n x_i x_i^t = \begin{pmatrix} n & \sum_{i=1}^n (z_i - \bar{z}) \\ \sum_{i=1}^n (z_i - \bar{z}) & \sum_{i=1}^n (z_i - \bar{z})^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & S_{zz} \end{pmatrix}. \end{aligned}$$

Somit hat X^tX den vollen Rang falls mindestens zwei $\{z_i\}$ verschieden sind. In dieser Situation ist nach Korollar §1.2.5 der gKQS eindeutig gegeben durch $\hat{\beta} = (X^tX)^{-1}X^tY = (\bar{Y}, S_{zz}^{-1}S_{zY})^t$ und somit sind $\hat{a} = \bar{Y} - \hat{b}\bar{z}$ und $\hat{b} = S_{zz}^{-1}S_{zY}$ die gKQS von a und b . \square

§1.2.8 **Varianzanalyse mit einem Faktor** (§1.1.5 (b) fortgesetzt). Wir bestimmen im Folgenden die gKQS der unbekannt Parameter μ_1, \dots, μ_p . Bezeichnet $\bar{Y}_{j\bullet} := q^{-1}\sum_{k=1}^q Y_{jk}$, $j = 1, \dots, p$, dann gilt $X^tY = (q\bar{Y}_{1\bullet}, \dots, q\bar{Y}_{p\bullet})$ und $X^tX = q \text{Id}_p$. Offensichtlich hat X^tX den vollen Rang so dass $\hat{\beta} = (X^tX)^{-1}X^tY = (\bar{Y}_{1\bullet}, \dots, \bar{Y}_{p\bullet})^t$ nach Korollar §1.2.5 der eindeutige gKQS von $\beta = (\mu_1, \dots, \mu_p)^t$ ist. \square

1.3 Der Satz von Gauß-Markov

§1.3.1 **Satz.** Besitzt die Designmatrix X den Rang $\text{rg}[X] = p$, so gelten im gewöhnlichen linearen Modell $Y \odot \{\mathfrak{L}(X\beta, \sigma^2 \text{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ die folgenden Aussagen:

(a) Der gKQS $\hat{\beta} = (X^tX)^{-1}X^tY$ ist ein erwartungstreuer Schätzer von β (d.h. $\mathbb{E}(\hat{\beta}) = \beta$).

(b) (**Satz von Gauß-Markov**) Unter allen Schätzern des abgeleiteten linearen Parameters $\gamma = \langle \beta, v \rangle$ für ein $v \in \mathbb{R}^p$, die linear (in den Daten Y) und für alle $\beta \in \mathbb{R}^p$ erwartungstreu sind, besitzt der lineare und erwartungstreue Schätzer $\hat{\gamma} = \langle \hat{\beta}, v \rangle$ eine minimale Varianz, nämlich $\text{Var}(\hat{\gamma}) = \sigma^2 \|X(X^t X)^{-1} v\|^2$.

(c) Bezeichnet $R := Y - X\hat{\beta}$ den Residuenvektor, so ist die geeignet normalisierte Stichprobenvarianz $\hat{\sigma}^2 := \frac{1}{n-p} \|R\|^2 = \frac{1}{n-p} \|Y - X\hat{\beta}\|^2$ ein erwartungstreuer Schätzer von σ^2 . \square

Beweis von Satz §1.3.1. in der Vorlesung. \square

§1.3.2 **Bemerkung.** (a) Der Schätzer $\hat{\gamma}$ im Satz von Gauß-Markov wird bester linearer erwartungstreuer Schätzer (**BLUE** für best linear unbiased estimator) genannt. Verzichtet man auf die Linearität oder Erwartungstreue des Schätzers, so gibt es im Allgemeinen bessere Schätzer im Sinne des mittleren quadratischen Fehlers, zumindest für ausgewählte unbekannte Parameter β bzw. γ . Ein einfacher linearer aber nicht erwartungstreuer Schätzer ist $\tilde{\gamma} = 0$. Offensichtlich gilt für seinen MSE $\mathbb{E}(\tilde{\gamma} - \gamma)^2 = \gamma^2$, so dass für alle unbekannt Parameter in einer hinreichend kleine Umgebung um die Null, der MSE von $\tilde{\gamma}$ strikt kleiner als der MSE des BLUE $\hat{\gamma}$ ist.

(b) Häufig sind wir nicht am MSE für eine Parameterschätzung im zu Grunde liegenden Modell interessiert, sondern an dem Vorhersagefehler $\|X\hat{\beta} - X\beta\|^2$. In der Situation einer gewöhnlichen linearen Regression entspricht dies der quadrierten Differenz der vorhergesagten und wahren Werte an den Designpunkten. Der Koordinaten des Vektors $\hat{Y} = X\hat{\beta}$ werden angepasste Werte (*fitted values*) genannt. Für den mittleren Vorhersagefehler (MPE für *mean prediction error*) prüft man nun leicht dass

$$\mathbb{E}\|X\hat{\beta} - X\beta\|^2 = \mathbb{E}\|\Pi_{\mathcal{R}(X)}Y - \Pi_{\mathcal{R}(X)}X\beta\|^2 = \mathbb{E}\|\Pi_{\mathcal{R}(X)}\varepsilon\|^2 = \sigma^2 p.$$

Insbesondere wächst der Vorhersagefehler linear in der Dimension p des Parameterraumes.

(c) Eine entsprechende Aussage des Satzes von Gauß-Markov gilt auch im linearen Modell $Y \odot \{\mathcal{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma\}$ (Übung!). \square

1.4 Die multivariate Normalverteilung

Nicht degenerierte multivariate Normalverteilungen können direkt über ihre Dichte definiert werden. Eine Normalverteilung heißt degeneriert, falls ihre Kovarianzmatrix nicht strikt positiv definit ist (nicht vollen Rang hat). In der Vorlesung werden wir auch Zufallsvariablen mit degenerierten Normalverteilungen betrachten. Beispiele für solche Zufallsvariablen sind Projektionen von nicht degenerierten normalverteilten Zufallsvariablen auf lineare Teilräume. Dies ist etwa der Fall für $X\hat{\beta}$ im Falle einer deterministischen Designmatrix und unabhängigen normalverteilten Fehlern. Dies wird in der nächsten Sektion behandelt.

§1.4.1 **Lemma.** Sei $X \in \mathbb{R}^p$ eine ZV mit $\mathbb{E}\|X\|^2 < \infty$. Für alle $b \in \mathbb{R}^q$ und $A \in \mathbb{R}^{q \times p}$ ist dann $Y = AX + b \in \mathbb{R}^q$ eine ZV mit $\mathbb{E}\|Y\|^2 < \infty$. Bezeichnen wir weiterhin mit $\mu := \mathbb{E}(X) \in \mathbb{R}^p$ und $\Sigma := \text{Cov}(X) \in \mathbb{R}^{p \times p}$ den Erwartungswert und die Kovarianzmatrix von X , dann gilt $\mathbb{E}(Y) = A\mu + b$ und $\text{Cov}(Y) = A\Sigma A^t$. \square

Beweis von Lemma §1.4.1. in der Vorlesung. \square

§1.4.2 **Satz (Cramér-Wold).** Die Verteilung einer ZV $X \in \mathbb{R}^p$ ist vollständig festgelegt durch die eindimensionalen Verteilungen der linear Formen $\langle X, c \rangle$ für alle $c \in \mathbb{R}^p$. □

Beweis von Satz §1.4.2. zum Beispiel unter Zuhilfenahme von multivariaten charakteristischen Funktionen, z.Bsp. Theorem 15.55 in Klenke [2008]. □

§1.4.3 **Korollar.** Die Koordinaten einer ZV $X \in \mathbb{R}^p$ sind genau dann unabhängig und identisch (standardnormal) $\mathfrak{N}(0, 1)$ -verteilt, wenn für alle $c \in \mathbb{R}^p$ die reellwertige ZV $\langle X, c \rangle$ eine $\mathfrak{N}(0, \langle c, c \rangle)$ -Verteilung besitzt, d.h. $\langle X, c \rangle$ ist stetig verteilt mit Dichte

$$f(x) = \frac{1}{(2\pi\langle c, c \rangle)^{1/2}} \exp\left(-\frac{x^2}{2\langle c, c \rangle}\right), \quad x \in \mathbb{R}. \quad \square$$

Beweis von Korollar §1.4.3. in der Vorlesung. □

§1.4.4 **Definition.** Ein zufälliger Vektor $X \in \mathbb{R}^p$ mit Erwartungswertvektor $\mu \in \mathbb{R}^p$ und Kovarianzmatrix $\Sigma := \text{Cov}(X) \in \mathbb{R}^{p \times p}$ besitzt eine **multivariate Normalverteilung**, falls für alle $c \in \mathbb{R}^p$ die reellwertige ZV $\langle X, c \rangle$ eine $\mathfrak{N}(\langle \mu, c \rangle, \langle \Sigma c, c \rangle)$ -Verteilung besitzt. Wir schreiben dann $X \sim \mathfrak{N}(\mu, \Sigma)$. Die Verteilung $\mathfrak{N}(0, \text{Id}_p) = \mathfrak{N}^{\otimes p}(0, 1)$ heißt insbesondere (*p-dimensionale Standardnormalverteilung*). □

§1.4.5 **Proposition.** Seien $X \sim \mathfrak{N}(0, \text{Id}_p)$ und $Y \sim \mathfrak{N}(0, \text{Id}_q)$, dann gelten die folgenden Aussagen

- (a) Falls $A \in \mathbb{R}^{m \times p}$ und $B \in \mathbb{R}^{m \times q}$ mit $AA^t = BB^t$ gilt, dann sind die ZV'en $AX \in \mathbb{R}^m$ und $BY \in \mathbb{R}^m$ identisch verteilt.
- (b) Falls $U \in \mathbb{R}^{m \times p}$ eine partielle Isometrie ist, dann gilt $UX \sim \mathfrak{N}(0, \Pi_{\mathcal{R}(U)})$.
- (c) Falls $A \in \mathbb{R}^{p \times m}$ und $B \in \mathbb{R}^{p \times q}$ mit $A^t B = 0$. Dann sind $\Pi_{\mathcal{R}(A)} X \sim \mathfrak{N}(0, \Pi_{\mathcal{R}(A)})$ und $\Pi_{\mathcal{R}(B)} X \sim \mathfrak{N}(0, \Pi_{\mathcal{R}(B)})$ unabhängig. □

Beweis von Proposition §1.4.5. in der Vorlesung. □

§1.4.6 **Korollar.** Sei $X \sim \mathfrak{N}(\mu, \Sigma)$, dann gelten die folgenden Aussagen:

- (a) Die *i*-te Koordinate von X ist $\mathfrak{N}(\mu_i, \Sigma_{ii})$ -verteilt.
- (b) Die Koordinaten von X sind genau dann unabhängig, wenn sie unkorreliert sind.
- (c) Für $A \in \mathbb{R}^{p \times q}$ und $b \in \mathbb{R}^q$ gilt $Y = AX + b \sim \mathfrak{N}(A\mu + \beta, A\Sigma A^t)$.
- (d) Ist Σ strikt positiv-definit, dann ist X stetig verteilt mit Lebesgue-Dichte

$$f(x) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} \exp\left\{-\frac{1}{2}\langle \Sigma^{-1}(x - \mu), (x - \mu) \rangle\right\}, \quad x \in \mathbb{R}^p. \quad \square$$

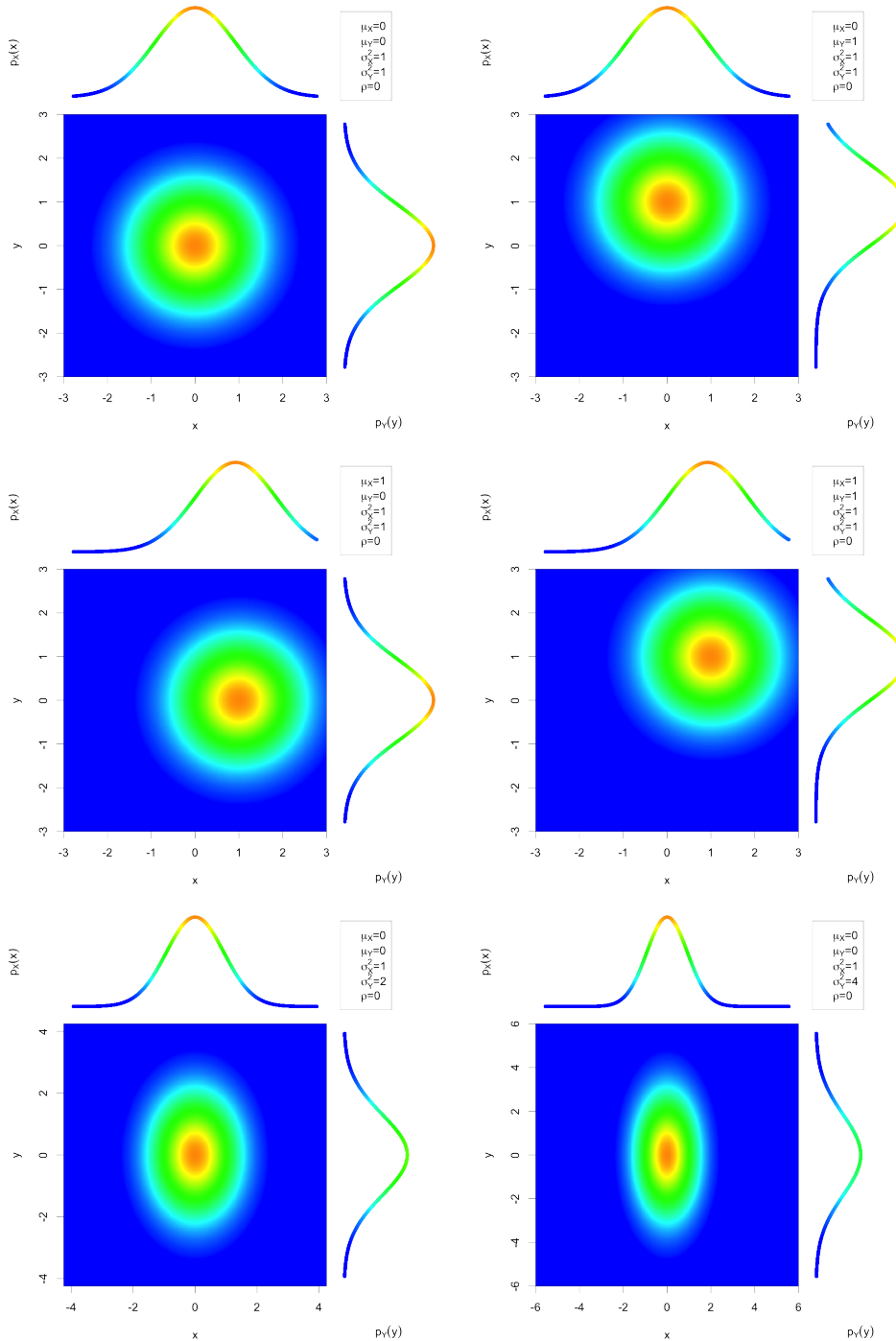
Beweis von Korollar §1.4.6. (Übung). □

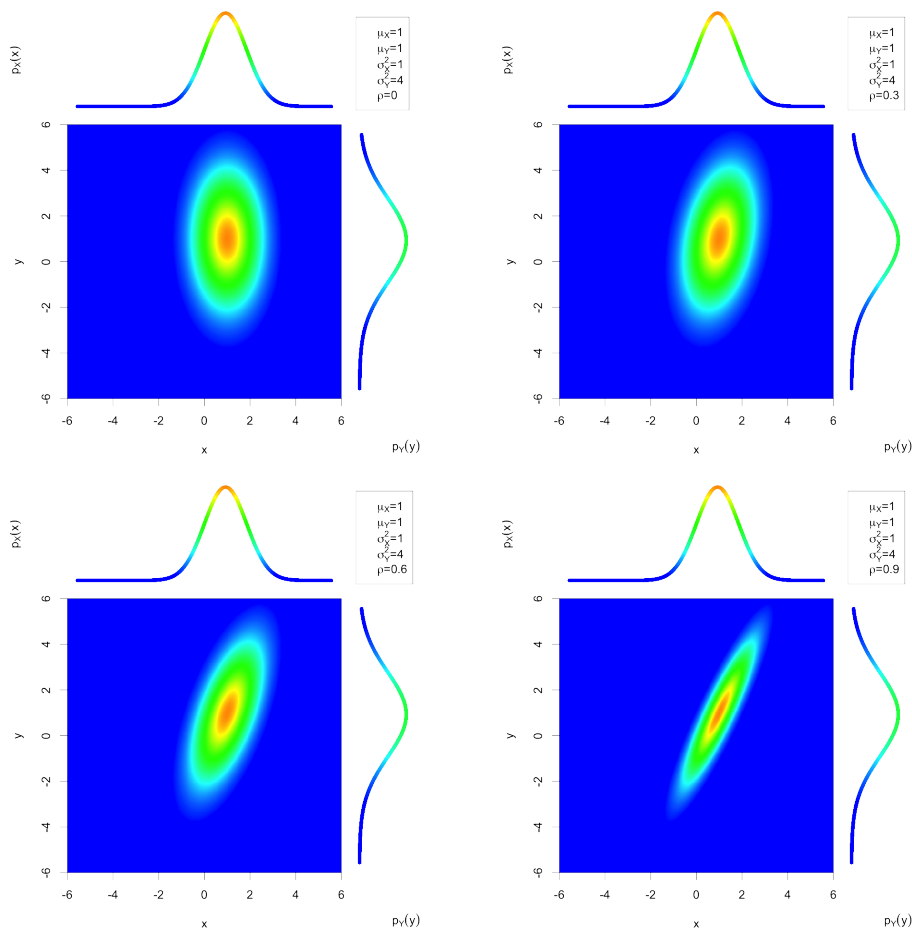
§1.4.7 **Beispiel.** Seien X und Y reellwertige ZV'en mit $\mathbb{E}(X^2) < \infty$ und $\mathbb{E}(Y^2) < \infty$. Setze $\mu_X := \mathbb{E}(X)$, $\mu_Y := \mathbb{E}(Y)$, $\sigma_X^2 := \text{Var}(X)$, $\sigma_Y^2 := \text{Var}(Y)$ und den Korrelationskoeffizienten $\rho := \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$. Der zufällige Vektor (X, Y) besitzt eine **bivariate Normalverteilung**, falls für alle Konstanten $a, b \in \mathbb{R}$ die ZV $aX + bY$ eine $\mathfrak{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2 +$

$2ab\rho\sigma_x\sigma_y$)-Verteilung besitzt. Die gemeinsame Dichte ist gegeben durch

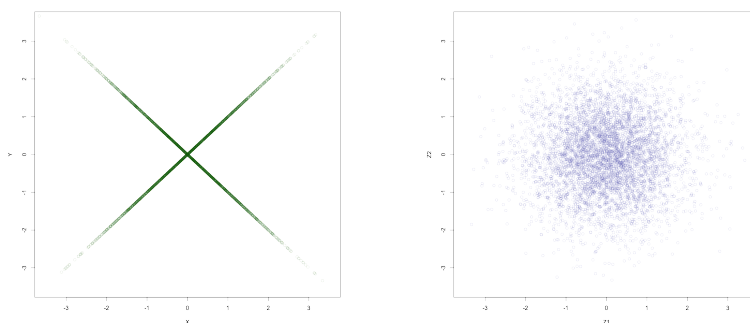
$$p(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left(-\frac{(x-\mu_X)^2}{2(1-\rho^2)\sigma_X^2}\right) \times \exp\left(\frac{2\rho(x-\mu_X)(y-\mu_Y)}{2(1-\rho^2)\sigma_X\sigma_Y}\right) \times \exp\left(-\frac{(y-\mu_Y)^2}{2(1-\rho^2)\sigma_Y^2}\right), \quad x, y \in \mathbb{R}.$$

Die nächsten Graphiken stellen die gemeinsame sowie die marginalen Dichten für verschiedene Werte der Parameter dar:





Besitzt (X, Y) eine *bivariate Normalverteilung* so gilt offensichtlich $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ und $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. Sind X und Y weiterhin unkorreliert, d.h. $\rho = 0$, dann sind X und Y unabhängig und es gilt $aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$. Insbesondere sind die folgenden beiden Aussagen äquivalent: (i) $X \sim \mathcal{N}(0, \sigma^2)$ und $Y \sim \mathcal{N}(0, \sigma^2)$ sind unabhängig; (ii) $X + Y \sim \mathcal{N}(0, 2\sigma^2)$ und $X - Y \sim \mathcal{N}(0, 2\sigma^2)$ sind unabhängig. (Warum?) Es ist natürlich möglich, dass $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ und $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ unkorreliert sind, aber der Vektor (X, Y) besitzt keine bivariate Normalverteilung. Betrachte dazu zwei unabhängige ZV'en X und V , wobei $X \sim \mathcal{N}(0, 1)$ und V ist eine Rademacher-ZV, d.h. $V \in \{-1, 1\}$ mit $P(V = -1) = 1/2 = P(V = 1)$. Es ist nun leicht zu zeigen, dass die ZV'en $Y := VX$ und X unkorreliert sind und dass $Y \sim \mathcal{N}(0, 1)$ (Übung!). Die ZV'en X und Y sind somit standardnormalverteilt und unkorreliert, aber ihre gemeinsame Verteilung ist keine Normalverteilung (warum?). Die nächsten Graphiken zeigen 5000 Realisierungen von (X, Y) (in grün) und zum Vergleich 5000 Realisierungen einer bivariaten Standardnormalverteilung.

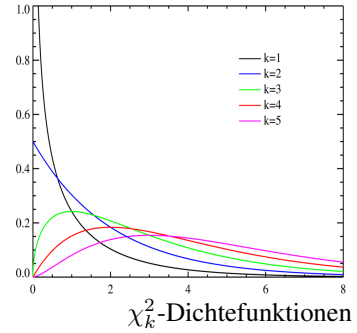


□

§1.4.8 **Definition.** Sei $(Z_1, \dots, Z_k)^t \sim \mathfrak{N}(0, \sigma^2 \text{Id}_k)$. Die Verteilung der ZV

$$Q := \sum_{i=1}^k Z_i^2$$

heißt **(zentrale) χ^2 -Verteilung** mit k *Freiheitsgraden*. Wir schreiben $Q \sim \chi_k^2$. Für $\alpha \in (0, 1)$ bezeichnen wir weiterhin den Wert $\chi_{k,\alpha}^2 \in \mathbb{R}$ als α -Quantil einer (zentralen) χ^2 -Verteilung mit k Freiheitsgraden, falls $P(Q \leq \chi_{k,\alpha}^2) = \alpha$. Für $\delta \in \mathbb{R}$ heißt die Verteilung der ZV



$$Q := (Z_1 + \delta)^2 + \sum_{i=2}^k Z_i^2$$

nichtzentrale χ^2 -Verteilung mit k *Freiheitsgraden* und *Nichtzentralitätsparameter* δ^2 . Wir schreiben $Q \sim \chi_k^2(\delta^2)$ sowie $\chi_{k,\alpha}^2(\delta^2) \in \mathbb{R}$ für das α -Quantil einer nichtzentralen χ^2 -Verteilung mit k Freiheitsgraden und Nichtzentralitätsparameter δ^2 , d.h. $P(Q \leq \chi_{k,\alpha}^2(\delta^2)) = \alpha$. □

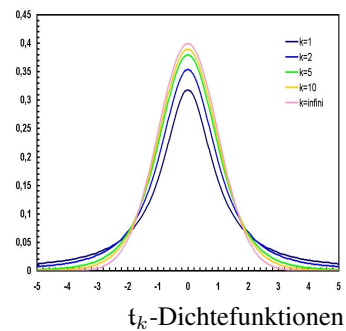
§1.4.9 **Korollar.** Sei $Q \sim \chi_k^2$ und $W \sim \chi_k^2(\delta^2)$, dann gilt $\mathbb{E}(Q) = k$, $\text{Var}(Q) = 2k$ und $\mathbb{E}(W) = \delta^2 + k$. Für $Z \sim \mathfrak{N}(0, \sigma^2 \text{Id}_m)$, $v \in \mathbb{R}^m$ und $A \in \mathbb{R}^{m \times p}$ mit $\text{rg}(A) = p$ gelten außerdem: (i) $\sigma^{-2} \|\Pi_{\mathcal{R}(A)} Z\|^2 \sim \chi_p^2$ und (ii) $\|Z/\sigma + v\|^2 \sim \chi_m^2(\|v\|^2)$. □

Beweis von Korollar §1.4.9. Übung. □

§1.4.10 **Definition.** Sei $(Z_0, Z_1, \dots, Z_k)^t \sim \mathfrak{N}(0, \sigma^2 \text{Id}_{k+1})$. Die Verteilung der ZV

$$T := \frac{Z_0}{\sqrt{\frac{1}{k} \sum_{i=1}^k Z_i^2}}$$

heißt **(Student-) t-Verteilung** mit k *Freiheitsgraden*. Wir schreiben: $T \sim t_k$ und bezeichnen mit $t_{k,\alpha}$ das α -Quantil einer Student-t-Verteilung mit k -Freiheitsgraden, d.h. $P(T \leq t_{k,\alpha}) = \alpha$.



§1.4.11 **Bemerkung.** Die Student-t-Verteilung mit einem ($k = 1$) Freiheitsgrad entspricht gerade der Cauchy-Verteilung und für $k \rightarrow \infty$ konvergiert sie schwach gegen die Standardnormalverteilung (Slutsky-Lemma). Für jedes $k \in \mathbb{N}$ besitzt die t_k -Verteilung endliche Momente nur bis zur Ordnung $p < k$ (sie ist heavy-tailed). Insbesondere, ist $T \sim t_k$ so gilt $\mathbb{E}(T) = 0$ für $k > 1$, sowie $\text{Var}(T) = k/(k - 2)$ für $k > 2$. □

§1.4.12 **Definition.** Sei $(Z_1, \dots, Z_{m+k})^t \sim \mathfrak{N}(0, \sigma^2 \text{Id}_{m+k})$.

Die Verteilung der ZV

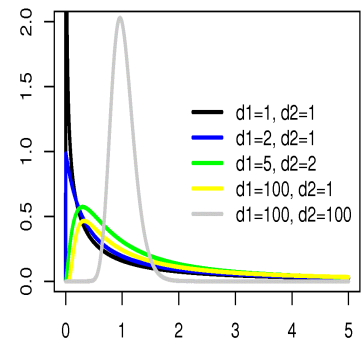
$$F := \frac{\frac{1}{m} \sum_{i=1}^m Z_i^2}{\frac{1}{k} \sum_{i=m+1}^{m+k} Z_i^2}$$

heißt **zentrale (Fisher-) \mathfrak{F} -Verteilung** mit m und k *Freiheitsgraden*. Wir schreiben: $F \sim \mathfrak{F}_{m,k}$ und bezeichnen mit $\mathfrak{F}_{m,k,\alpha}$ das α -Quantil einer zentralen Fisher- \mathfrak{F} -Verteilung mit m und k Freiheitsgraden, d.h. $P(F \leq \mathfrak{F}_{m,k,\alpha}) = \alpha$. Für $\delta \in \mathbb{R}$ heißt die Verteilung der ZV

$$F := \frac{\frac{1}{m} \{(Z_1 + \delta)^2 + \sum_{i=2}^m Z_i^2\}}{\frac{1}{k} \sum_{i=m+1}^{m+k} Z_i^2}$$

nichtzentrale (Fisher-) \mathfrak{F} -Verteilung mit m und k *Freiheitsgraden* und *Nichtzentralitätsparameter* δ^2 . Wir schreiben $F \sim \mathfrak{F}_{m,k}(\delta^2)$ sowie $\mathfrak{F}_{m,k,\alpha}(\delta^2) \in \mathbb{R}$ für das α -Quantil einer nichtzentralen \mathfrak{F} -Verteilung mit m und k Freiheitsgraden und Nichtzentralitätsparameter δ^2 , d.h. $P(F \leq \mathfrak{F}_{m,k,\alpha}(\delta^2)) = \alpha$. \square

§1.4.13 **Bemerkung.** Sei $F \sim \mathfrak{F}_{m,k}$ mit $k > 1$, dann ist F^{-1} eine $\mathfrak{F}_{k,m}$ -verteilte ZV. Für $T \sim t_k$ ist T^2 eine $\mathfrak{F}_{1,k}$ -verteilte ZV. Weiterhin sei $F_k \sim \mathfrak{F}_{m,k}$, $k \in \mathbb{N}$, dann konvergiert die Folge von ZV'en $(mF_k)_{k \geq 1}$ für $k \rightarrow \infty$ in Verteilung gegen ein χ_m^2 -verteilte ZV. \square



$\mathfrak{F}_{d1,d2}$ -Dichtefunktionen

1.5 Das normale lineare Modell

§1.5.1 **Definition.** Ein **normales lineares Modell** bezeichnet ein lineares Modell in dem der zu erklärende zufällige Vektor eine multivariate Normalverteilung besitzt. Beobachtet wird eine Realisierung von Y und die Designmatrix X und wir schreiben abkürzend $Y \odot \{\mathfrak{N}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$. In einem **gewöhnlichen normalen linearen Modell** gilt weiterhin $\Sigma = \sigma^2 \text{Id}_n$ für ein Fehlerniveau $\sigma > 0$. Im gewöhnlichen Fall sind die Koordinaten des zentrierten Fehlervektors $\varepsilon := Y - X\beta$ unabhängig und identisch $\mathfrak{N}(0, \sigma^2)$ -verteilt, d.h. $\varepsilon/\sigma \sim \mathfrak{N}^{\otimes n}(0, 1)$. \square

§1.5.2 **Satz.** *Besitzt die Designmatrix X den Rang $\text{rg}[X] = p$, so gelten im gewöhnlichen normalen linearen Modell $Y \odot \{\mathfrak{N}(X\beta, \sigma^2 \text{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ die folgenden Aussagen:*

(a) *Der gKQS ist normalverteilt:*

$$\hat{\beta} \sim \mathfrak{N}(\beta, \sigma^2(X^t X)^{-1}).$$

(b) *Die Stichprobenvarianz $\hat{\sigma}^2 = \frac{1}{n-p} \|Y - X\hat{\beta}\|^2$ ist nach geeigneter Normalisierung χ^2 -verteilt mit $n - p$ Freiheitsgraden:*

$$(n - p) \hat{\sigma}^2 / \sigma^2 \sim \chi_{n-p}^2.$$

(c) *Der gKQS $\hat{\beta}$ und die Stichprobenvarianz $\hat{\sigma}^2$ sind unabhängig.*

(d) Der zentrierte und geeignet normalisierte gKQS $\widehat{\beta}$ hat eine t -Verteilung mit $n - p$ Freiheitsgraden: für $v \in \mathbb{R}^p$

$$\frac{\langle \widehat{\beta} - \beta, v \rangle}{\widehat{\sigma} | \langle (X^t X)^{-1} v, v \rangle |^{1/2}} \sim t_{n-p}.$$

(e) Der Vorhersagefehler $\|X(\beta - \widehat{\beta})\|^2$ ist nach geeigneter Normalisierung \mathfrak{F} -verteilt mit p und $n - p$ Freiheitsgraden:

$$\frac{\|X(\widehat{\beta} - \beta)\|^2}{p\widehat{\sigma}^2} \sim \mathfrak{F}_{p, n-p}.$$

Beweis von Satz §1.5.2. in der Vorlesung. □

§1.5.3 **Korollar.** Unter den Annahmen und den Notationen des Satzes §1.5.2 gelten folgende Konfidenzaussagen für gegebenes $\alpha \in (0, 1)$:

(a) **Konfidenzbereich für β :** Bezeichnet $\mathfrak{F}_{p, n-p, 1-\alpha}$ das $(1 - \alpha)$ -Quantil einer \mathfrak{F} -Verteilung mit p und $n - p$ Freiheitsgraden, so ist

$$C_\alpha = \{ \beta \in \mathbb{R}^p : \|X(\widehat{\beta} - \beta)\|^2 < p\widehat{\sigma}^2 \mathfrak{F}_{p, n-p, 1-\alpha} \}$$

ein Konfidenzellipsoid zum Niveau $1 - \alpha$ für β .

(b) **Konfidenzbereich für $\langle \beta, v \rangle$:** Bezeichnet $t_{\alpha} := t_{n-p, 1-\alpha/2} = -t_{n-p, \alpha/2}$ das $(1 - \alpha/2)$ -Quantil einer t -Verteilung mit $n - p$ Freiheitsgraden, so ist

$$I_{v, \alpha} = [\langle \widehat{\beta}, v \rangle - t_{\alpha} \widehat{\sigma} | \langle (X^t X)^{-1} v, v \rangle |^{1/2}, \langle \widehat{\beta}, v \rangle + t_{\alpha} \widehat{\sigma} | \langle (X^t X)^{-1} v, v \rangle |^{1/2}]$$

ein Konfidenzintervall zum Niveau $1 - \alpha$ für $\langle \beta, v \rangle$.

§1.5.4 **Beispiel** (§1.1.5 (a) fortgesetzt). In einem *normalen Lokations-Skalen-Modell*

$Y \odot \{ \mathfrak{N}^n(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0 \}$ ist

$$I_{v, \alpha} = [\bar{Y} - t_{n-1, 1-\alpha/2} n^{-1/2} \widehat{\sigma}, \bar{Y} + t_{n-1, 1-\alpha/2} n^{-1/2} \widehat{\sigma}]$$

mit $\widehat{\sigma}^2 = \frac{1}{n-1} \|Y - \bar{Y} \mathbb{1}_n\|^2$ ein Konfidenzintervall zum Niveau $1 - \alpha$ für den unbekannt Parameter μ . Dies folgt direkt aus Korollar §1.5.3 (b) mit $p = 1$, $v = 1$ und $\gamma = \mu$. □

§1.5.5 **Korollar.** Unter den Annahmen und den Notationen des Satzes §1.5.2 kann für ein $r \in \mathbb{R}$ die **lineare Hypothese** $H_0 : \langle \beta, v \rangle = r$ gegen die Alternativen (a) $H_A : \langle \beta, v \rangle > r$; (b) $H_A : \langle \beta, v \rangle < r$ sowie (c) $H_A : \langle \beta, v \rangle \neq r$ mit Hilfe der Teststatistik $T := \frac{\langle \widehat{\beta}, v \rangle - r}{\widehat{\sigma} | \langle (X^t X)^{-1} v, v \rangle |^{1/2}}$ und den Entscheidungsregeln

(a) lehne die Hypothese H_0 ab, falls $T \geq t_{n-p, 1-\alpha}$;

(b) lehne die Hypothese H_0 ab, falls $T \leq -t_{n-p, 1-\alpha}$;

(c) lehne die Hypothese H_0 ab, falls $|T| \geq t_{n-p, 1-\alpha/2}$;

unter Einhaltung des vorgegebenen Niveau $\alpha \in (0, 1)$ getestet werden.

§1.5.6 **Beispiel** (§1.5.4 fortgesetzt). In einem *normalen Lokations-Skalen-Modell*

$Y \odot \{ \mathfrak{N}^n(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0 \}$ kann die **Hypothese** $H_0 : \mu = \mu_o$ gegen die Alternativen (a) $H_A : \mu > \mu_o$; (b) $H_A : \mu < \mu_o$ sowie (c) $H_A : \mu \neq \mu_o$ mit Hilfe der Entscheidungsregeln

- (a) lehne die Hypothese H_0 ab, falls $\bar{Y} - \mu_o > t_{n-1, 1-\alpha} n^{-1/2} \hat{\sigma}$;
 (b) lehne die Hypothese H_0 ab, falls $\bar{Y} - \mu_o < t_{n-1, 1-\alpha} n^{-1/2} \hat{\sigma}$;
 (c) lehne die Hypothese H_0 ab, falls $|\bar{Y} - \mu_o| > t_{n-1, 1-\alpha/2} n^{-1/2} \hat{\sigma}$;

unter Einhaltung des vorgegebenen Niveau $\alpha \in (0, 1)$ getestet werden. \square

1.6 Asymptotische Theorie

Wir untersuchen nun die Verteilung des Kleinste-Quadrate-Schätzers im Grenzfall, in dem die Anzahl der Beobachtungen gegen unendlich geht. Dazu sei $(Y_n)_{n \in \mathbb{N}}$ eine Folge von zufälligen Zielgrößen und $(x_n)_{n \in \mathbb{N}}$ eine Folge von erklärenden Effekten. Wir nehmen an, dass für alle $n \geq n_0$ der Zusammenhang zwischen dem zufälligen Vektor $Y_{(n)} := (Y_1, \dots, Y_n)^t$ und der Designmatrix $X_{(n)} = (x_1, \dots, x_n)^t$ adäquat durch ein gewöhnliches lineares Modell beschrieben ist, d.h. $Y_{(n)} \odot \{\mathcal{L}(X_{(n)}\beta, \sigma^2 \text{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$.

§1.6.1 **Satz.** Sei $Y_{(n)} \odot \{\mathcal{L}(X_{(n)}\beta, \sigma^2 \text{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ mit $\text{rg}[X_{(n)}] = p$ für alle $n \geq n_0$. Gelten die folgenden drei Bedingungen:

- (i) $\{Y_n - x_n^t \beta, n \in \mathbb{N}\}$ sind unabhängige und identisch verteilte (u.i.v.) ZV'en.
 (ii) Für den kleinsten Eigenwert $\lambda_{(n)}$ der Matrix $X_{(n)}^t X_{(n)}$ gilt $\lim_{n \rightarrow \infty} \lambda_{(n)} = \infty$.
 (iii) Für die Diagonalelemente der Matrix $P_{(n)} := X_{(n)}(X_{(n)}^t X_{(n)})^{-1} X_{(n)}^t$ gilt $\lim_{n \rightarrow \infty} \max_{j=1, \dots, n} [P_{(n)}]_{jj} = 0$.

Dann ist der Kleinste-Quadrate-Schätzer $\hat{\beta}_{(n)} := (X_{(n)}^t X_{(n)})^{-1} X_{(n)}^t Y_{(n)}$ konsistent für β und

$$\frac{1}{\sigma} (X_{(n)}^t X_{(n)})^{1/2} (\hat{\beta}_{(n)} - \beta) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \text{Id}_k)$$

(konvergiert in Verteilung gegen eine k -dimensionale Standardnormalverteilung) und weiterhin gilt für $v \in \mathbb{R}^p$

$$\frac{\langle \hat{\beta}_{(n)} - \beta, v \rangle}{\sigma | \langle (X_{(n)}^t X_{(n)})^{-1} v, v \rangle |^{1/2}} \xrightarrow{\mathcal{L}} \mathfrak{N}(0, 1).$$

Gilt zusätzlich $\mathbb{E}(Y_1 - x_1^t \beta)^4 < \infty$, dann ist $\hat{\sigma}^2 = \frac{1}{n-p} \|Y_{(n)} - X \hat{\beta}_{(n)}\|^2$ ein konsistenter Schätzer für σ^2 . \square

Beweis von Satz §1.6.1. in der Vorlesung. \square

§1.6.2 **Bemerkung.** Die Bedingung §1.6.1 (ii) besagt, dass man mit wachsendem n immer mehr Information bekommt. Weiterhin dominiert kein Vektor von Effekten x_j die anderen unter der Bedingung §1.6.1 (iii). \square

§1.6.3 **Korollar.** Unter den Annahmen und den Notationen des Satzes §1.6.1 gilt folgende asymptotische Konfidenzaussage für gegebenes $\alpha \in (0, 1)$. Bezeichnet $z_{1-\alpha/2}$ das $(1 - \alpha/2)$ -Quantil einer $\mathfrak{N}(0, 1)$ -Verteilung, so ist

$$I_{v, \alpha} = \left[\langle \hat{\beta}_{(n)}, v \rangle - z_{1-\alpha/2} \hat{\sigma} | \langle (X_{(n)}^t X_{(n)})^{-1} v, v \rangle |^{1/2}, \langle \hat{\beta}_{(n)}, v \rangle + z_{1-\alpha/2} \hat{\sigma} | \langle (X_{(n)}^t X_{(n)})^{-1} v, v \rangle |^{1/2} \right]$$

ein Konfidenzintervall zum asymptotischen Niveau $1 - \alpha$ für $\langle \beta, v \rangle$.

§1.6.4 **Beispiel** (§1.1.5 (a) fortgesetzt). Sei $Y_{(n)} \odot \{\mathcal{L}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$ durch ein *Lokations-Skalen-Modell* mit u.i.v. Koordinaten adäquat beschrieben, dann ist

$$I_{v,\alpha} = [\bar{Y}_{(n)} - z_{1-\alpha/2} n^{-1/2} \hat{\sigma}, \bar{Y}_{(n)} + z_{1-\alpha/2} n^{-1/2} \hat{\sigma}]$$

mit $\bar{Y}_{(n)} = \frac{1}{n} \sum_{i=1}^n Y_i$ ein Konfidenzintervall zum asymptotischen Niveau $1 - \alpha$ für den unbekannt Parameter μ . Dies folgt direkt aus Korollar §1.6.3 mit $v = 1$. \square

§1.6.5 **Korollar**. Unter den Annahmen und den Notationen des Satzes §1.6.1 kann für ein $r \in \mathbb{R}$ die *lineare Hypothese* $H_0 : \langle \beta, v \rangle = r$ gegen die Alternativen (a) $H_A : \langle \beta, v \rangle > r$; (b) $H_A : \langle \beta, v \rangle < r$ sowie (c) $H_A : \langle \beta, v \rangle \neq r$ mit Hilfe der Teststatistik $T := \frac{\langle \hat{\beta}, v \rangle - r}{\hat{\sigma} | \langle (X^t X)^{-1} v, v \rangle |^{1/2}}$ und den Entscheidungsregeln

(a) lehne die Hypothese H_0 ab, falls $T \geq z_{1-\alpha}$;

(b) lehne die Hypothese H_0 ab, falls $T \leq -z_{1-\alpha}$;

(c) lehne die Hypothese H_0 ab, falls $|T| \geq z_{1-\alpha/2}$;

unter Einhaltung des vorgegebenen asymptotischen Niveau $\alpha \in (0, 1)$ getestet werden.

§1.6.6 **Beispiel** (§1.6.4 fortgesetzt). Sei $Y \odot \{\mathcal{L}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}\}$ durch ein *Lokations-Skalen-Modell* mit u.i.v. Koordinaten adäquat beschrieben, dann kann die *Hypothese* $H_0 : \mu = \mu_0$ gegen die Alternativen (a) $H_A : \mu > \mu_0$; (b) $H_A : \mu < \mu_0$ sowie (c) $H_A : \mu \neq \mu_0$ mit Hilfe der Entscheidungsregeln

(a) lehne die Hypothese H_0 ab, falls $\bar{Y}_{(n)} - \mu_0 > z_{1-\alpha} n^{-1/2} \sigma$;

(b) lehne die Hypothese H_0 ab, falls $\bar{Y}_{(n)} - \mu_0 < z_{1-\alpha} n^{-1/2} \sigma$;

(c) lehne die Hypothese H_0 ab, falls $|\bar{Y}_{(n)} - \mu_0| > z_{1-\alpha/2} n^{-1/2} \sigma$;

unter Einhaltung des vorgegebenen asymptotischen Niveau $\alpha \in (0, 1)$ getestet werden. \square

1.7 Residuenanalyse

Wir nehmen im Folgenden an, dass der Zusammenhang zwischen der Zielgröße Y und der Designmatrix durch ein gewöhnliches lineares Modell $Y \odot \{\mathcal{L}(X\beta, \sigma^2 \text{Id}_n)\}$ adäquat dargestellt ist. Bezeichnen wir mit \bar{Y} das arithmetische Mittel der Beobachtung, so ist die totale Quadratsumme $\|Y - \bar{Y} \mathbb{1}_n\|^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$ (SST für *total sum of squares*) ein Maß der Variabilität der Realisierungen der Zielgrößen. Wir wollen nun untersuchen in wie weit diese Variabilität durch die Variabilität der angepassten Schätzwerte $\hat{Y} = X\hat{\beta}$ oder der Residuen $Y - \hat{Y}$ erklärt wird. Eine einfache Zerlegung der totalen Quadratsumme in eine Quadratsumme der Regression bzgl. der angepassten Werte (SSR für *regression sum of squares*) und eine Quadratsumme der Residuen (SSE für *error sum of squares*) ergibt

$$SST := \|Y - \bar{Y} \mathbb{1}_n\|^2 = \|\hat{Y} - \bar{Y} \mathbb{1}_n\|^2 + \|Y - \hat{Y}\|^2 =: SSR + SSE.$$

Offensichtlich, spricht ein im Verhältnis zum SSR kleiner Wert des SSE für eine gute Anpassung des linearen Modells. Betrachten wir den standardisierten Quotienten

$$F = \frac{\frac{1}{p} SSR}{\frac{1}{n-p} SSE},$$

so sprechen große Werte von F für eine gute Anpassung des linearen Modells. Nehmen wir zusätzlich an, dass die Beobachtung Y normalverteilt ist, so vergleichen wir die Anpassung in dem linearen Modell $Y \odot \{\mathfrak{N}(X\beta, \sigma^2 \text{Id}_n)\}$ mit der in einem Lokations-Skalen-Modell $Y \odot \{\mathfrak{N}^{\otimes n}(\mu, \sigma^2)\}$. Unter der Annahme, dass $\mathbb{1} \in \mathcal{R}(X)$ gilt, hat F^* eine \mathfrak{F} -Verteilung mit $(p, n - p)$ Freiheitsgraden.

Alternativ können wir des Verhältnis zwischen der totalen Variabilität und der Variabilität der Schätzwerte betrachten:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Der Wert R^2 wird *Bestimmtheitsmaß* genannt und entspricht im Fall $k = 1$ dem Quadrat des *empirischen Korrelationskoeffizienten*

$$\rho = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Bezeichnet $\hat{\beta}_{-i}$ den gewöhnliche Kleinste-Quadrate-Schätzer ohne die i -te Koordinate der Beobachtung Y , dann gilt

$$\hat{\beta}_{-i} - \hat{\beta} = -\frac{\hat{Y}_i - Y_i}{1 - [X(X^t X)^{-1} X]_{ii}} (X^t X)^{-1} x_i.$$

Wir sehen also, dass der Einfluss der i -ten Beobachtung sowohl vom i -ten Residuum als auch vom Diagonalelement $[X(X^t X)^{-1} X]_{ii}$, seinem *Leverage-Score*, abhängt. Um einflussreiche Beobachtungen zu entdecken, plottet man daher oft die Residuen R_i gegen die $[X(X^t X)^{-1} X]_{ii}$. Basierend auf der Differenz der geschätzten Parameter ist die *Cook-Distanz* definiert durch

$$\frac{1}{\hat{\sigma}^2} \|\hat{\beta}_{-i} - \hat{\beta}\|_{X^t X} = \frac{1}{k \hat{\sigma}^2} \frac{(\hat{Y}_i - Y_i)^2}{1 - [X(X^t X)^{-1} X]_{ii}} \frac{[X(X^t X)^{-1} X]_{ii}}{1 - [X(X^t X)^{-1} X]_{ii}}.$$

Sie ist eine einfache Funktion von $[X(X^t X)^{-1} X]_{ii}$ sowie dem Quadrat des studentisierten Residuums $(\hat{Y}_i - Y_i) / \sqrt{\hat{\sigma}^2 (1 - [X(X^t X)^{-1} X]_{ii})}$ welche Student-t-verteilt ist unter einer Normalverteilungsannahme. Sie wird häufig als diagnostisches Hilfsmittel verwendet. Diejenigen Beobachtungen, bei denen die Cook-Distanz deutlich größer ist als beim Rest, sollte besonders betrachtet, bzw. in der Analyse weggelassen werden. Analog erhält man als Änderung beim Hinzufügen einer Beobachtung Y_{n+1} zum Effekt x_{n+1} :

$$\frac{1}{1 + x_{n+1}^t (X^t X)^{-1} x_{n+1}} (X^t X)^{-1} x_{n+1} (Y_{n+1} - x_{n+1}^t \hat{\beta}).$$

Durch Hinzufügen einer einzigen Beobachtung kann somit der Kleinste-Quadrate-Schätzer beliebig verändert werden.

Literaturverzeichnis

A. Klenke. *Probability theory. A comprehensive course*. London: Springer, 2008.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New York., 1998.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL <http://www.R-project.org/>.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.