Hermann Singer

Multivariate Statistik

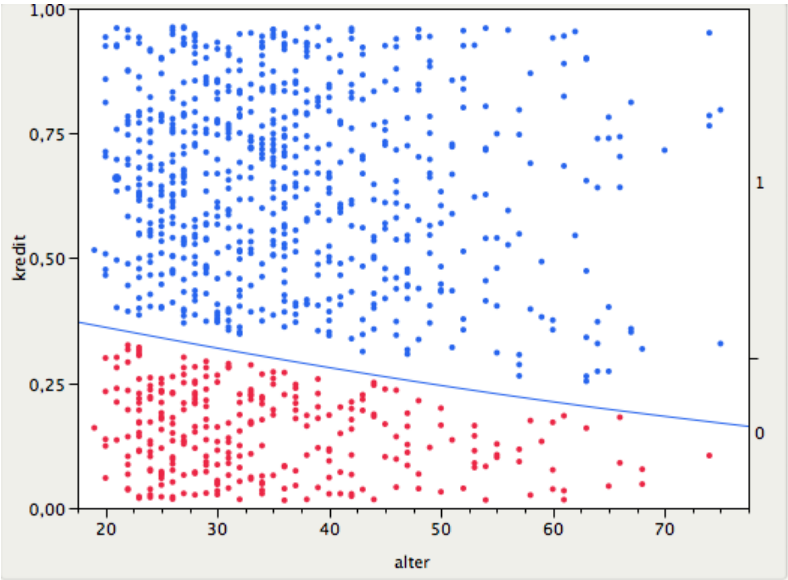

Abbildung 1.2: Logistische Regression: Geschätzte Wahrscheinlichkeit für schlechte und gute Kredite (rot/blau) als Funktion des Alters. Mit steigendem Alter sinkt die Wahrscheinlichkeit, dass der Kredit ausfällt, also $p(\texttt{Kredit} = \textit{schlecht}|\texttt{Alter})$.

Beispiel 1.2 (OECD-Daten)

Ein interessanter Datensatz, der auf OECD-Erhebungen beruht[2], soll im folgenden diskutiert werden. Er ist auf der Kurs-CD enthalten (als Excel-, SPSS- und JMP-Datei), kann jedoch auch im Internet als Excel-Tabelle gefunden werden (siehe `http://oecdbetterlifeindex.org/`). Die Excel-Tabelle kann in SPSS importiert und anschließend als SPSS-Datensatz (.sav) wieder gespeichert werden (siehe Abb. 1.3).

```
SPSS/Datei/Öffnen/Daten/Format Excel auswählen
```

Abb. 1.4 zeigt die sogenannte Daten- und Variablenansicht des Datensatzes `BetterLifeIndex.sav`. Der Datensatz wird mit Hilfe des Menüs

```
SPSS/Datei/Öffnen/Daten
```

[2]Organisation for Economic Cooperation and Development

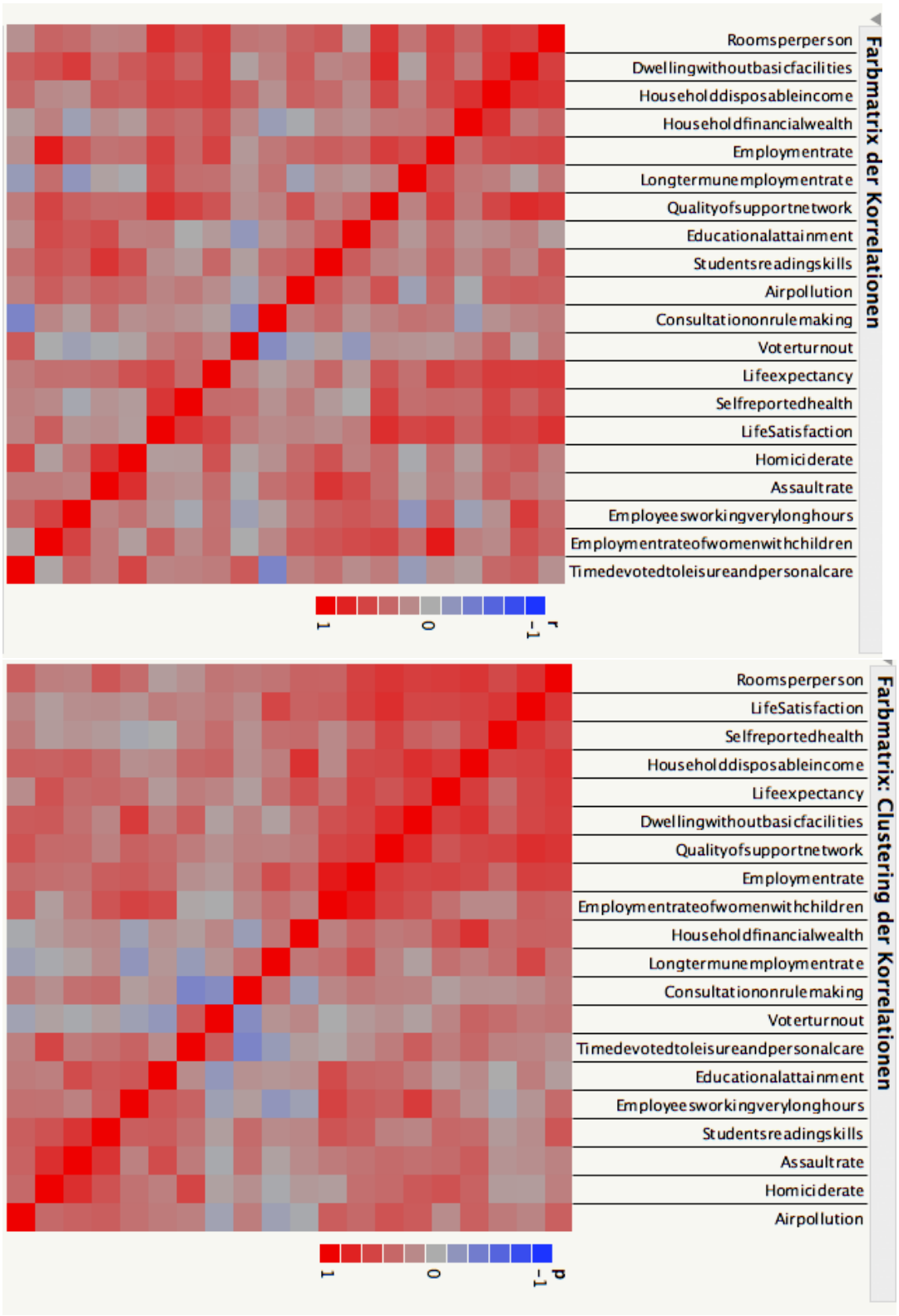

Abbildung 1.9: SAS/JMP: Farbmatrix der Korrelationen und Cluster (diagonales Ordnen).

Asymmetrische Verfahren $Y = f(X)$		
	UV $= X$	
AV $= Y$	diskret	stetig
diskret	Kreuztabellen, log-lineare Modelle kategoriale Regression	kategoriale Regression Diskriminanz-Analyse
stetig	Varianz-Analyse	Regressions-Analyse

allgemeines lineares Modell

Sind die abhängigen Variablen stetig und hat man gemischte stetige und diskrete unbhängige Variablen, so spricht man auch vom allgemeinen linearen Modell. Nimmt man bei der Varianz-Analyse (diskrete UV) noch stetige Kovariablen (d.h. weitere UV) hinzu, so ergibt sich das Modell der Kovarianzanalyse.

Kovarianzanalyse

Clusteranalyse

Verfahren, bei denen Objekte (Zeilen der Datenmatrix) anhand der Spalten (Variablen) gruppiert werden, entstammen dem Bereich der Clusteranalyse.

Faktorenanalyse

Hat man eine große Zahl korrelierter Variablen, so kann eine Dimensionsreduktion auf wenige latente Faktoren angestrebt werden (Faktorenanalyse).

Strukturgleichungs-Modellierung

Auch sind Kombinationen von Regressions- und Faktorenanalyse möglich. Dies wird als Strukturgleichungs-Modellierung bezeichnet.

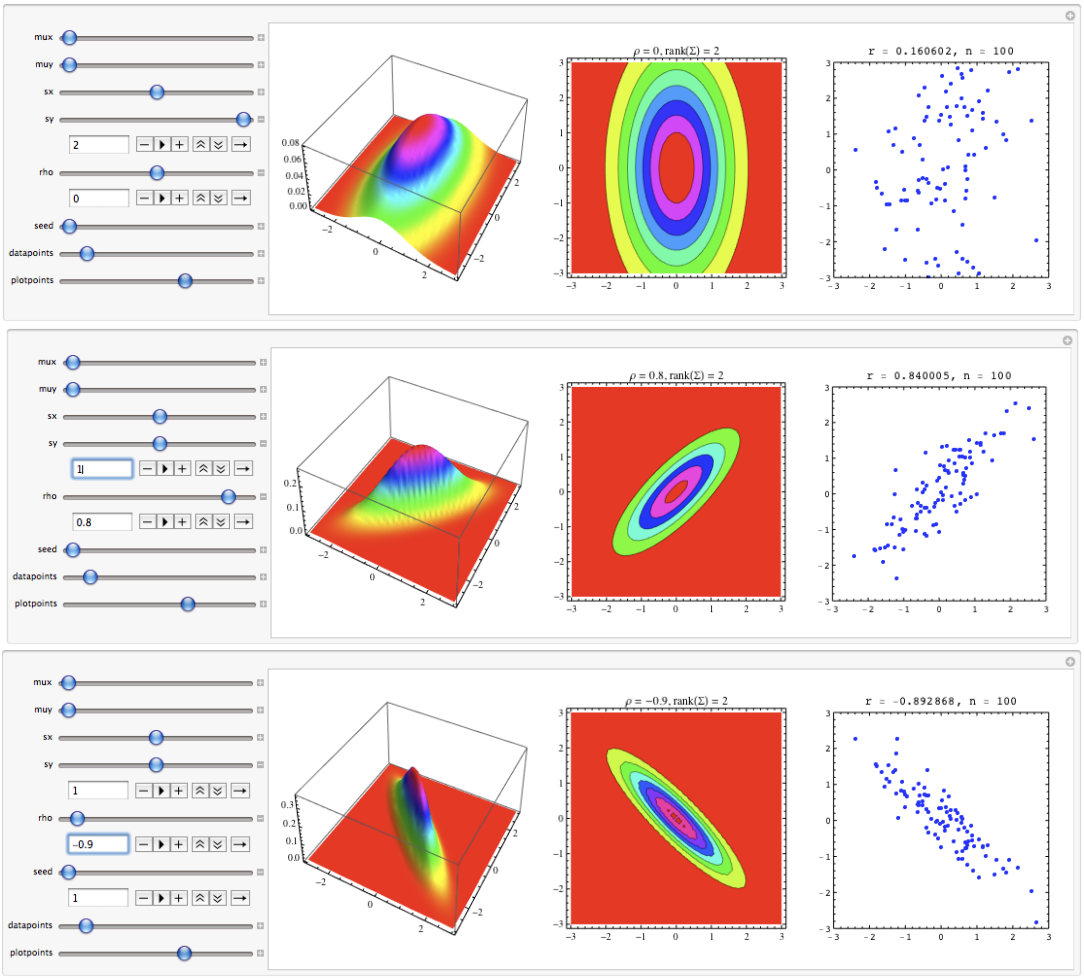

Abbildung 2.1: Bivariate Normalverteilungsdichte.
Obere Zeile: $\rho_{xy} = 0$, $\sigma_x = 1$, $\sigma_y = 2$. Mittlere Zeile: $\rho_{xy} = 0.8$ $\sigma_x = 1$, $\sigma_y = 1$. Untere Zeile: $\rho_{xy} = -0.9$ $\sigma_x = 1$, $\sigma_y = 1$.
Von Links: Regler, 3D-Graphik, Höhenlinien und simulierte Daten ($N = 100$).
`http://www.fernuni-hagen.de/ls_statistik/lehre/`

$[X_1, ..., X_p]' : p \times 1$ für den Zufallsvektor $\mathbf{x}$ (wird klein geschrieben, um eine Verwechslung mit der Matrix $\mathbf{X}$ zu vermeiden), so ist die p-variate Normalverteilungsdichte für $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ durch folgenden Ausdruck gegeben:

$$\phi(\mathbf{x}) = \det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{2.49}$$

Hierbei ist $\mathbf{x} = [x_1, ..., x_p]'$ ein p-Vektor und

$$\boldsymbol{\mu} = E[\mathbf{x}] = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \tag{2.50}$$

sowie

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \dots & \mathrm{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \dots & \mathrm{Cov}(X_p, X_p) \end{bmatrix} \tag{2.51}$$

sind die Parameter (Vektoren und Matrizen) der p-variaten Normalverteilung. Als Abkürzung kann man auch $\sigma_{ij} = \mathrm{Cov}(X_i, X_j), i, j = 1, ..., p$ schreiben. Hierbei ist $\sigma_{ii} = \sigma_i^2 = \mathrm{Var}(X_i)$ die Varianz und $\sigma_i = \sqrt{\sigma}_{ii}$ die Standardabweichung.

Der Korrelationskoeffizient zwischen den Variablen X_i und X_j, $i, j = 1, ..., p$,

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{2.52}$$

kann als Matrix $\mathbf{P}$ zusammengefasst werden. Schreibt man alle Standardabweichungen in eine Diagonalmatrix

$$\mathbf{D} = \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_p \end{bmatrix} = \mathrm{diag}(\sigma_1, ..., \sigma_p) \tag{2.53}$$

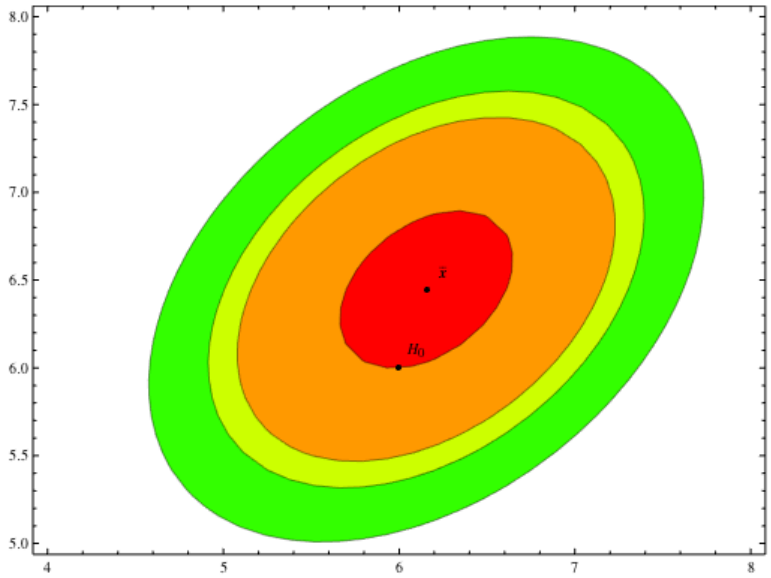

Abbildung 3.3: OECD-Daten. Unbekanntes $\boldsymbol{\Sigma}$. Konfidenz-Ellipsen zu den Niveaus $1-\alpha = 0.4, 0.9, 0.95, 0.99$. Außerdem ist die Nullhypothese H_0 : $\boldsymbol{\mu}_0 = [6,6]'$ eingezeichnet.

χ^2-Verteilung und der Hotelling-T^2-Verteilung ist in Abb. 3.4, unten) zu sehen. Die Quantile der Hotelling-T^2-Verteilung sind immer größer, da ja $\boldsymbol{\Sigma}$ nur geschätzt wurde (analog zur Normal- und t-Verteilung).

Wählt man als Nullhypothese $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = [7, 5.5]'$, so ergibt sich

$$\tilde{t}^2 = \tfrac{32}{66} \cdot 34 \, [-0.846, 0.947] \begin{bmatrix} 0.151 & -0.063 \\ -0.063 & 0.182 \end{bmatrix} \begin{bmatrix} -0.846 \\ 0.947 \end{bmatrix} = 6.135.$$

Damit muß H_0 auf dem 5%-Niveau abgelehnt werden (vgl. Abb. 3.6).

Der Stoff wird in Aufgabe 3.2 vertieft.

■

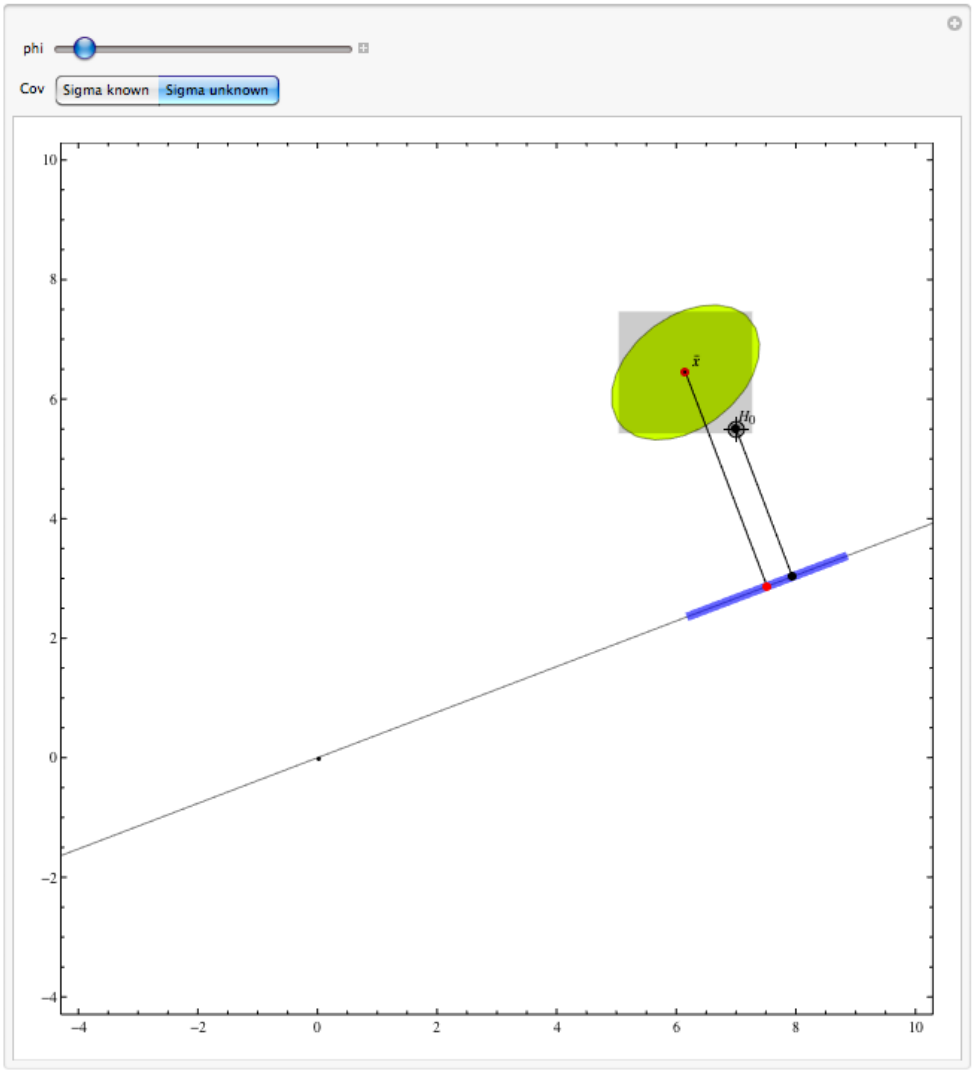

Abbildung 3.7: Applet für simultane Konfidenz-Intervalle.
`http://www.fernuni-hagen.de/ls_statistik/lehre/`

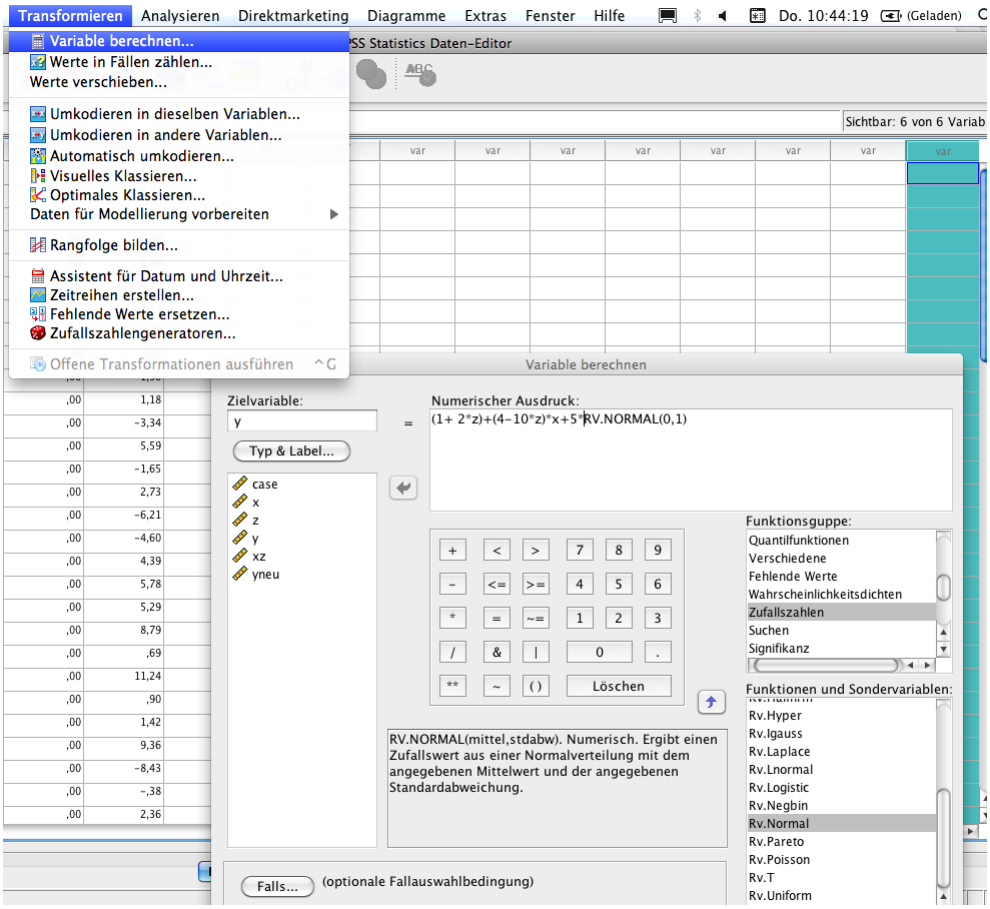

Abbildung 4.12: Berechnung der abhängigen Variablen y. Die wahren Parameterwerte sind $\beta_0 = 1, \beta_2 = 2, \beta_1 = 4, \beta_3 = -10, \sigma = 5$.

Das lineare Modell in Effekt-Kodierung lautet explizit:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1J} \\ Y_{21} \\ \vdots \\ Y_{2J} \\ \vdots \\ Y_{I-1,1} \\ \vdots \\ Y_{I-1,J} \\ Y_{I1} \\ \vdots \\ Y_{IJ} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ 1 & -1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & -1 & -1 & -1 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{I-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1J} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2J} \\ \vdots \\ \epsilon_{I-1,1} \\ \vdots \\ \epsilon_{I-1,J} \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{IJ} \end{bmatrix}$$

Der Parameter α_I, der in $\tilde{\boldsymbol{\mu}}$ nicht vorkommt, ergibt sich als $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$. Dies wird durch die negativen Einsen der letzten J Zeilen bewirkt.

Etwas kompakter kann man schreiben

$$\mathbf{y} = \left[\mathbf{1}_I \otimes \mathbf{1}_J, \begin{bmatrix} \mathbf{I}_{I-1} \\ -\mathbf{1}'_{I-1} \end{bmatrix} \otimes \mathbf{1}_J\right] \begin{bmatrix} \mu \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{\epsilon} \tag{5.56}$$

$$:= [\mathbf{X}_0, \mathbf{X}_\alpha] \begin{bmatrix} \mu \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{\epsilon}. \tag{5.57}$$

Die Abkürzung

Effektkodierung

$$x_{ii'}^{\alpha} = \begin{cases} 1, & i = i' < I \\ -1, & i = I \\ 0, & sonst \end{cases} \tag{5.58}$$

$i = 1, ..., I$, $i' = 1, ..., I-1$ bzw. als Matrix

$$\mathbf{x}^{\alpha} = \begin{bmatrix} \mathbf{I}_{I-1} \\ -\mathbf{1}'_{I-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -1 & \dots & -1 \end{bmatrix} : I \times (I-1) \tag{5.59}$$

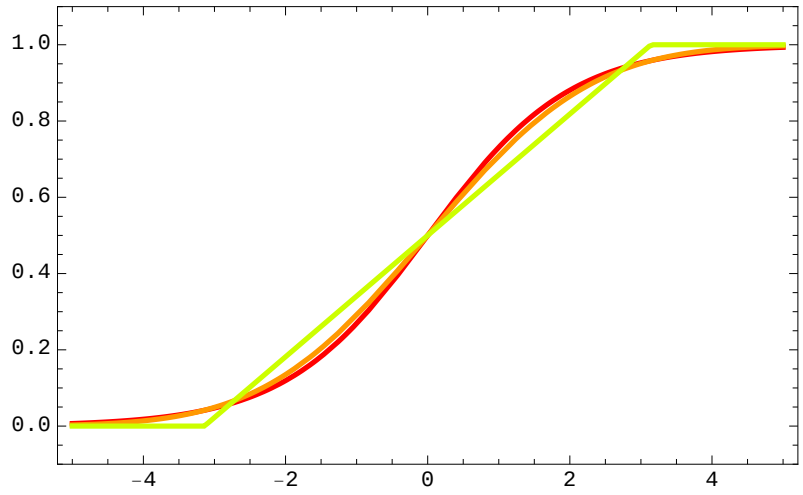

Abbildung 6.1: Responsefunktionen: Logistische (rot), Normalverteilung (orange), Gleichverteilung (grün). Die Varianzen wurden auf den Wert $\pi^2/3$ der logistischen Funktion adjustiert.

Probit-Modell

Probit-Modell

$$p(y = 1|\mathbf{x}) \quad = \quad \Phi(\mathbf{x}'\boldsymbol{\beta}). \tag{6.16}$$

Die unterschiedlichen Modell sind in Abb. 6.1 dargestellt. Zum besseren Vergleich wurden die Varianzen auf den Wert $\pi^2/3$ der logistischen Funktion adjustiert. Dies ist sinnvoll, da die Funktionen $h(\beta_0 + \beta_1 x) = \tilde{h}(\tilde{\beta}_0 + \tilde{\beta}_1 x)$ auf eine äquivalente Modellierung führen. Daher kann die Funktion verschoben und das Argument mit einem Faktor skaliert werden (vgl. Fahrmeir et al., 1996, S. 249). Die Unterschiede in den Funktionen sind recht gering, wobei die logistische Funktion im Gegensatz zur Normalverteilung leichter zu berechnen ist.

Generell muß die Response-Funktion zwischen 0 und 1 liegen, es ist nicht notwendig, daß es sich um eine kumulative Verteilungsfunktion handelt.

Man kann jedoch das binäre Regressions-Modell durch eine latente Variable $Y^* = \mathbf{x}'\boldsymbol{\beta}^* + \epsilon$ motivieren, die nicht direkt beobachtet werden

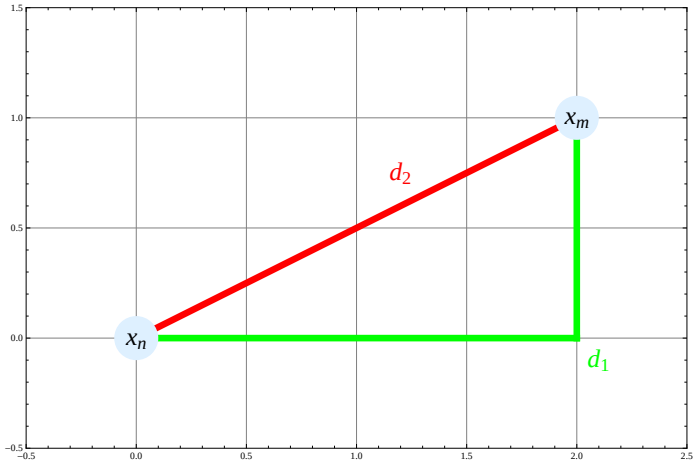

Abbildung 7.3: Vergleich von euklidischer Distanz d_2 und City-Block-Metrik d_1. Diese bleibt invariant, wenn andere kürzeste Wege entlang des Rasters genommen werden.

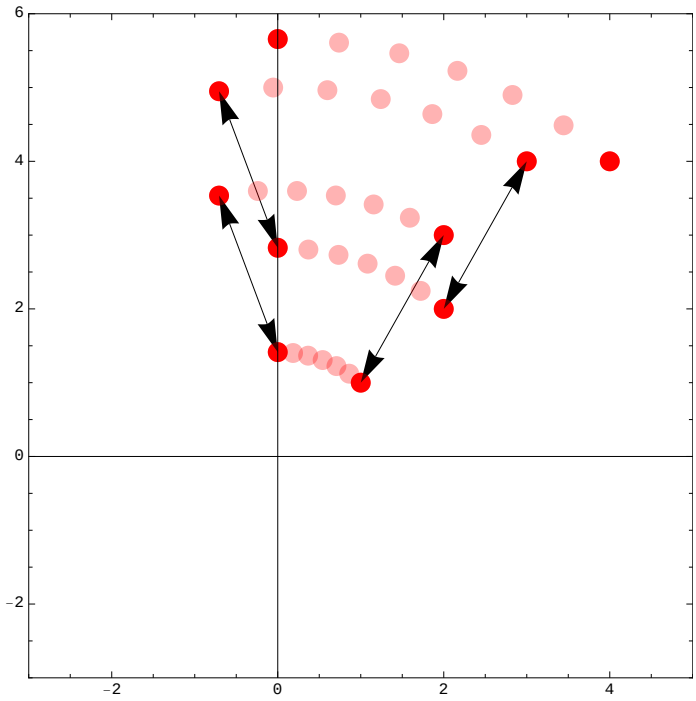

Abbildung 7.4: Daten und Abstände. Translationsinvarianz der Distanzen. Bei um ϕ rotierten Daten bleiben die Abstände invariant.

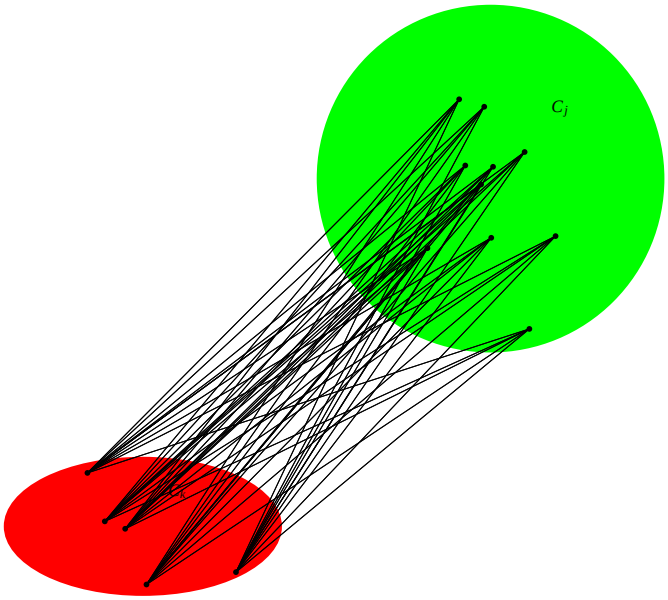

Abbildung 7.11: Abstand von 2 Klassen beim average-linkage-Verfahren.

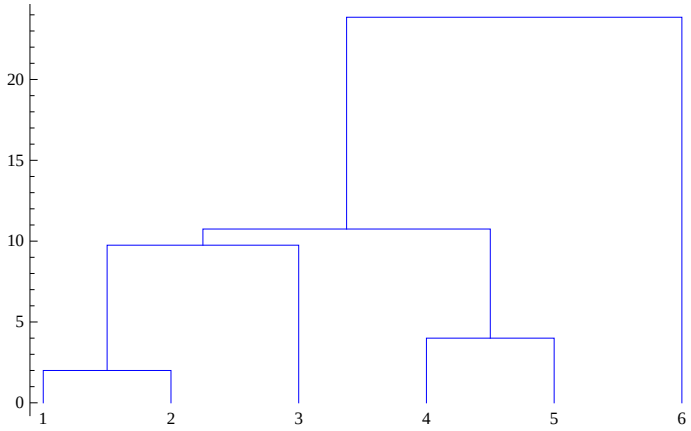

Abbildung 7.12: Dendrogramm beim average-linkage-Verfahren.

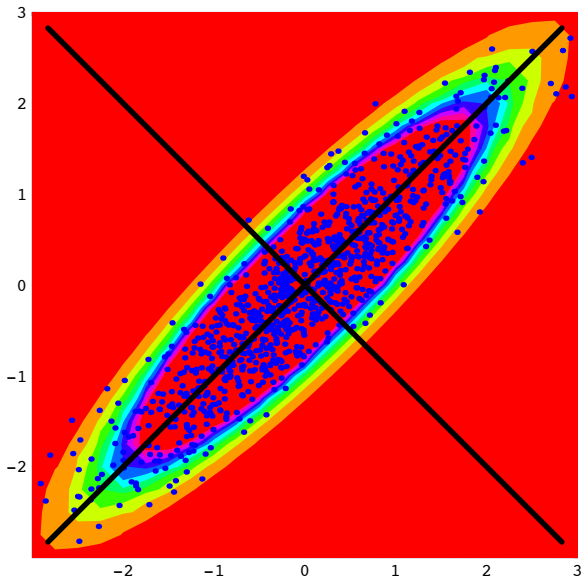

Abbildung 8.2: Simulierte normalverteilte Daten $\mathbf{x}_n, n = 1, ..., N = 1000$ mit Kovarianz-Matrix $\mathbf{R} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$. Die Hauptachsen zeigen in Richtung der Winkelhalbierenden.

Daher sind die gedrehten Koordinaten (Hauptkomponenten) y_1, y_2 unkorreliert.
Die quadratische Form (Ellipse) der Matrix $\mathbf{R}$

$$\mathbf{x}'\mathbf{R}\mathbf{x} = \sum_{ij} x_i \rho_{ij} x_j = x_1^2 + 2\rho x_1 x_2 + x_2^2 \tag{8.44}$$

ist diagonal im gedrehten System:

$$\begin{aligned} \mathbf{x}'\mathbf{R}\mathbf{x} &= \mathbf{x}'\mathbf{P}\mathbf{P}'\mathbf{R}\mathbf{P}\mathbf{P}'\mathbf{x} && (8.45)\\ &= \mathbf{y}'\mathbf{M}\mathbf{y} = \mu_1 y_1^2 + \mu_2 y_2^2 = (1+\rho)y_1^2 + (1-\rho)y_2^2. && (8.46) \end{aligned}$$

Die im Bild gezeigte Ellipse ist allerdings

$$\begin{aligned} \mathbf{x}'\mathbf{R}^{-1}\mathbf{x} &= \mathbf{y}'\mathbf{M}^{-1}\mathbf{y} && (8.47)\\ &= \frac{y_1^2}{\mu_1} + \frac{y_2^2}{\mu_2} && (8.48)\\ &= \frac{y_1^2}{1+\rho} + \frac{y_2^2}{1-\rho} && (8.49)\\ &= \frac{y_1^2}{1.9} + \frac{y_2^2}{0.1}, && (8.50) \end{aligned}$$