SBWL Tourismusanalyse und Freizeitmarketing

Vertiefungskurs 4: Multivariate Verfahren 2

Teil 1: Verallgemeinerte lineare Modelle

Achim Zeileis

Inhalt

- Einheit 1: Überblick
  - Motivation
  - Methoden
  - Datensätze
- Einheit 2: EDA
  - Bivariate explorative Datenanalyse (EDA)
  - Mosaikplot
  - Wiederholung: (multiple) lineare Regression

Inhalt

- Einheit 3: Schätztechniken
  - Kleinste-Quadrate-Schätzung
  - Maximum-Likelihood-Schätzung
- Einheit 4: (Ko-)Varianzanalyse
  - qualitative Erklärungsvariablen: Varianzanalyse
  - Kontraste
  - Kovarianzanalyse
- Einheit 5: Modellwahl
  - Inferenz
  - Informationskriterien

Inhalt

- Einheit 6: Binäre Variablen
  - Kontigenztafeln
  - Odds Ratio
  - Mosaikplots
- Einheit 7:
  - Logistische Regression
  - Poisson-Regression

Überblick: Motivation

Dieser Teil der LV beschäftigt sich mit der Modellierung des Zusammenhangs einer abhängigen Variablen von mehreren Variablen verschiedener Datenniveaus.

Ziel ist dabei die Erstellung eines guten Erklärungsmodells bzw. eines Prognosemodells für die abhängige Variable.

Die Wahl einer geeigneten Methode hängt dabei vor allem vom Skalenniveau der abhängigen Variablen ab:

- **qualitativ** (oder kategorial)
- **quantitativ** (oder numerisch, metrisch, kontinuierlich)

Überblick: Motivation

Insbesonders werden Modelle diskutiert für

- (approximativ) **normalverteilte** Beobachtungen, (d.h. etwa symmetrisch unimodal verteilt).
  Beispiel: Ausgaben in einem Urlaub.
  Modell: Normalverteilung.
- **binäre** Daten.
  Beispiel: Kaufentscheidungen.
  Modell: Binomialverteilung.
- **Zähldaten**.
  Beispiel: Mengenentscheidungen.
  Modell: Poissonverteilung.

Überblick: Motivation

Verteilungen:

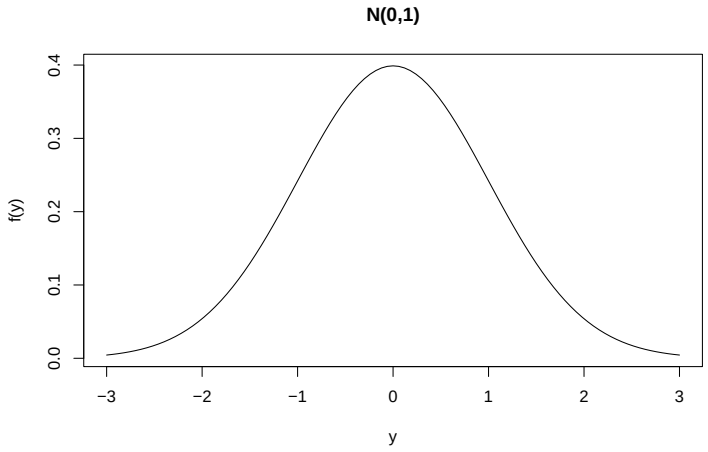
Normalverteilung:

$$f(y \mid \mu, \sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{(y-\mu)^2}{\sigma^2}\right)$$

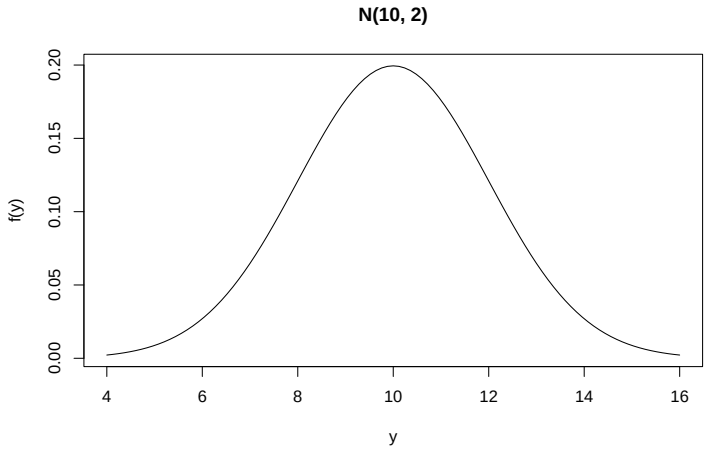
mit $-\infty \leq y \leq \infty$.

Die Normalverteilung ist ein Modell für metrische Daten, die unimodal sind und in etwa symmetrisch.

Überblick: Motivation



Überblick: Motivation



Überblick: Motivation

Binomialverteilung:

$$f(y \mid n, p) \quad = \quad \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$$

mit $y = 0, 1, \ldots, n$.

Die Binomialverteilung ist ein Modell für die Anzahl "Erfolge" bei n unabhängigen Versuchen und einer Erfolgswahrscheinlichkeit p.

Einfaches Beispiel: Anzahl Kopf bei n-maligem Münzwurf.

Überblick: Motivation

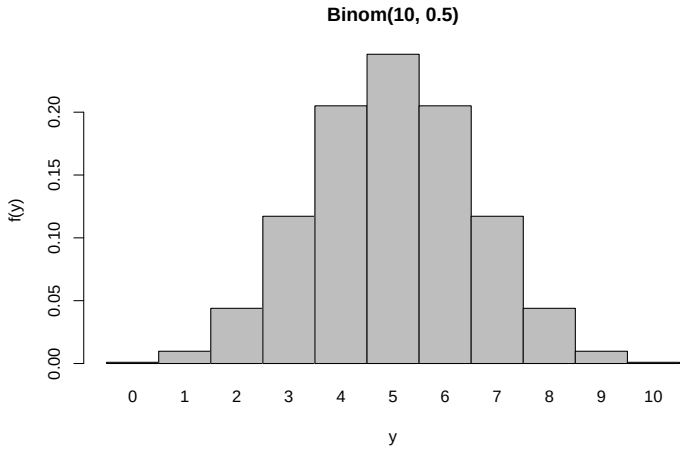
Besseres Beispiel: y Kunden einer Kundengruppe mit n Personen, die sich für ein bestimmtes Produkt entschieden haben.

Oft ist $n = 1$, d.h. der Kunde kauft das Produkt oder nicht. Dann:

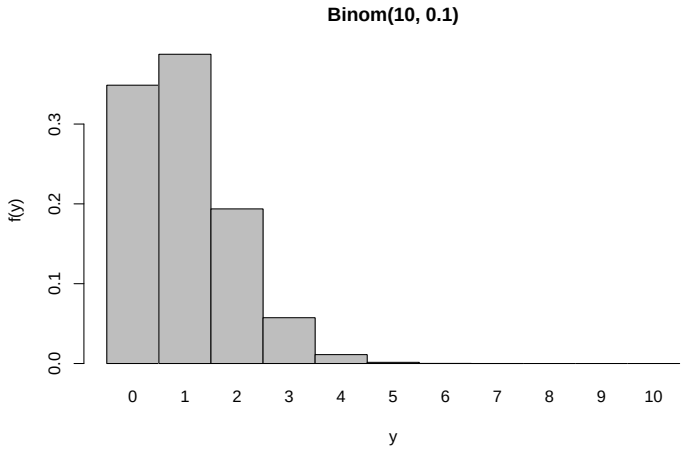
$$\begin{aligned} f(1) &= p, \\ f(0) &= 1 - p. \end{aligned}$$

Also ist p die Kaufswahrscheinlichkeit.

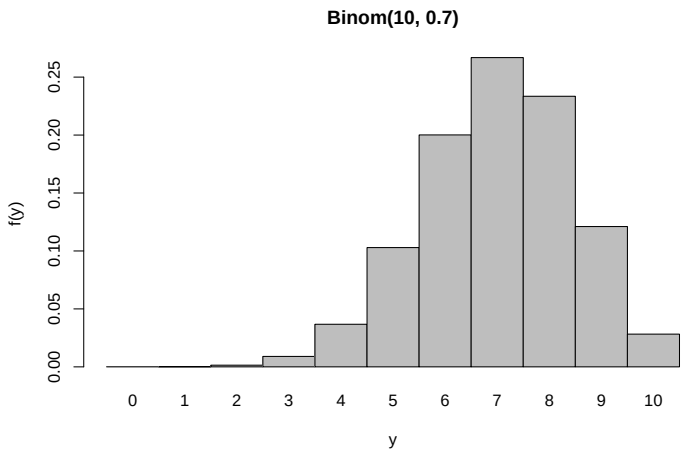
Überblick: Motivation



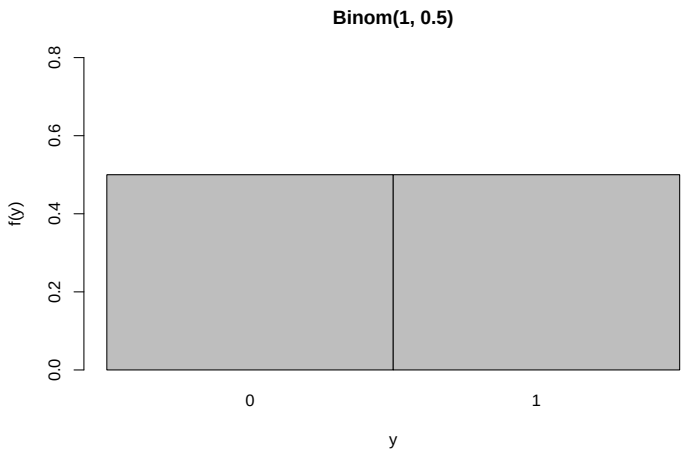
Überblick: Motivation



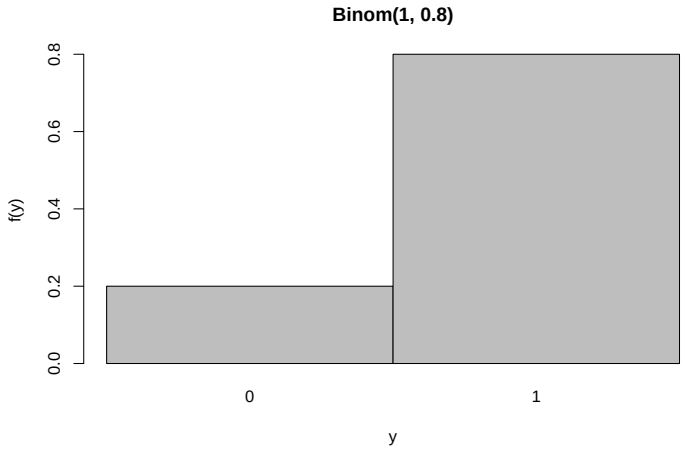
Überblick: Motivation



Überblick: Motivation



Überblick: Motivation



Überblick: Motivation

Poissonverteilung:

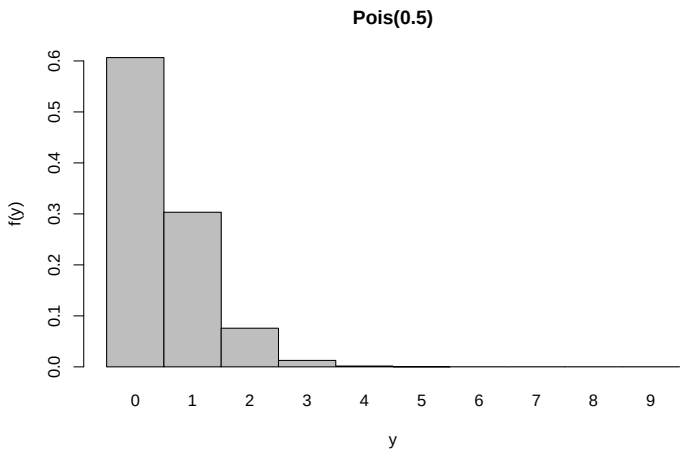
$$f(y \mid \lambda) \quad = \quad \frac{\lambda^y \cdot \exp(-\lambda)}{y!}$$

mit $y = 0, 1, 2, \ldots$.

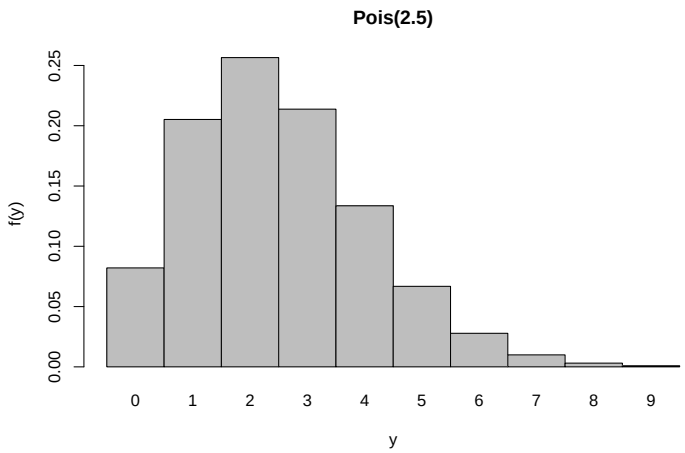
Die Poissonverteilungen ist ein Modell für Zähldaten.

Beispiel: Anzahl gekaufter Produkte bei einem Supermarktbesuch.

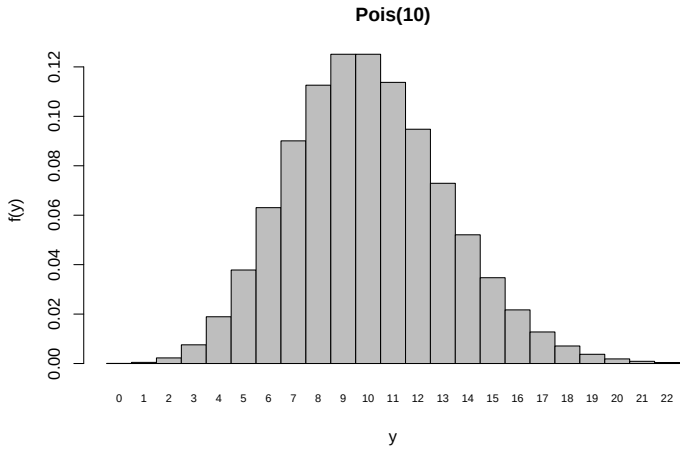
Überblick: Motivation



Überblick: Motivation



Überblick: Motivation



Überblick: Motivation

Grundlage für eine modellgestützte Analysen sollten aber auch immer deskriptive Analysen sein, da diese ein Gespür für die Daten vermittelt und das Verständnis und die Interpretation von statistischen Modellen erleichtert.

Ein kurzer Überblick über die Methoden zur deskriptiven Analyse von univariaten und bivariaten Datensätzen soll hier kurz gegeben werden.

Überblick: Methoden

1 quantitative Variable

Numerische Beschreibung: Mittelwert, Varianz, Standardabweichung, Fünf-Punkt-Zusammenfassung (Minimum, unteres Quartil, Median, oberes Quartil, Maximum)

Grafische Beschreibung: Histogramm, Boxplot.

1 qualitative Variable

Numerische Beschreibung: Häufigkeitstablle (absolute und relative Häufigkeiten).

Grafische Beschreibung: Balkendiagramm.

Überblick: Methoden

2 quantitative Variablen

Numerische Beschreibung: Korrelationskoeffizient.

Grafische Beschreibung: Streudiagramm.

2 qualitative Variablen

Numerische Beschreibung: Kontigenztafeln, Odds Ratio.

Grafische Beschreibung: Mosaikplot.

Überblick: Methoden

1 abhängige quantitative und 1 erklärende qualitative Variable

Numerische Beschreibung: Gruppierte Statistiken.

Grafische Beschreibung: parallele Boxplots.

1 abhängige qualitative und 1 erklärende quantitative Variable

Numerische Beschreibung: diskretisierte Kontigenztafeln.

Grafische Beschreibung: diskretisierter Mosaikplot.

Überblick: Methoden

Alle Modelle dieses Teils der LV gehören zu den verallgemeinerten linearen Regressionsmodellen (GLMs). Diese verallgemeinern das lineare Modell, wozu die einfache lineare Regression, multiple lineare Regression sowie die Varianzanalyse (ANOVA) gehören

Bei der linearen Regression wird eine metrische Variable durch ebenfalls metrische Variablen erklärt; bei der Varianzanalyse durch eine kategoriale Variable. Von Kovarianzanalyse spricht man, wenn es sowohl quantitative als auch qualitative Erklärungsvariablen gibt.

Das GLM verallgemeinert all diese Ideen zu Modellen, in denen die abhängige Variable nicht zwingend metrisch ist, sondern u.a. auch binär oder diskret.

Überblick: Daten

`BBBClub`

Der Bookbinder's Book Club ist ein amerikanischer Bücherclub, der 20,000 Kunden eine Brochure für das Buch "The Art History of Florence" hat. Von diesen haben 1806 Kunden dieses Buch daraufhin gekauft. Der BBB Club hat verschiedene Variablen dieser Kunden erhoben, um damit ein Prognosemodell für die Kaufentscheidung zu entwickeln. Einen Ausschnitt von 1,300 Beobachtungen ist verfügbar im Datensatz *BBBClub.rda* (bzw. *BBBClub.csv*) mit folgenden Variablen:

Überblick: Daten

- `choice` Hat der Kunde das Buch "The Art History of Florence" gekauft?
- `gender` Geschlecht.
- `amount` Gesamtsumme der Ausgaben beim BBB Club.
- `freq` Gesamtanzahl von Käufen beim BBB Club.
- `last` Monate seit dem letzten Kauf.
- `first` Monate seit dem ersten Kauf.
- `child` Anzahl gekaufter Kinderbücher.
- `youth` Anzahl gekaufter Jugendbücher.
- `cook` Anzahl gekaufter Kochbücher.
- `diy` Anzahl gekaufter Do-It-Yourself-Bücher.
- `art` Anzahl gekaufter Kunstbücher.

Überblick: Daten

`GSA`

In der Gästebefragung Österreich (Guest Survey Austria) wurden insgesamt 14,571 Touristen befragt, die Österreich in den Sommersaisons 1994 bzw. 1997 besucht haben. Zusätzlich zu bestimmten soziodemografischen Daten wie Alter, Geschlecht, Beruf usw. wurden die Touristen insbesonders nach bevorzugten Sommeraktivitäten gefragt (Tennis, Wandern, Theater, uva.) sowie nach ihren Motiven für die Auswahl ihres Urlaubsziels (Entspannung, Sport, Kultur, uva.).

Überblick: Daten

- `age` Alter in Jahren.
- `occupation` Beruf bzw. Beschäftigung.
- `income` Monatliches Haushaltseinkommen in AUS.
- `gender` Geschlecht.
- `country` Herkunftsland.
- `expenditure` Ausgaben pro Tag in AUS.
- `province` Zielbundesland.
- `accomodation` Unterkunftstyp.
- `year` Jahr des Besuches.
- `SA`$nn.xxx$ Sommer-Aktivität ausgeübt oder nicht?
- `M`$nn.xxx$ War das Motiv ein Grund für den Urlaub oder nicht?

Tutorium

Explorative Datenanalyse in R (*EDA.pdf*)

Übung

Aufgabe 1:

Betrachten Sie die die ersten 9 Variablen der Guest Survey Austria Daten (`GSA`). D.h. also *ohne* die `SA`nn`.`xxx und `M`nn`.`xxx Variablen.

Führen Sie eine univariate explorative Datenanalyse jeder der 9 Variablen durch.

Bei einigen der metrischen Variablen kann eine Transformation durch den Logarithmus `log` oder die Quadratwurzel `sqrt` hilfreich sein. Finden Sie heraus bei welchen.

Explorative Datenanalyse

Bivariate explorative Datenanalyse:

abh. vs. erkl.	**quantitativ**	**qualitativ**
quantitativ	Streudiagramm Korrelation	Boxplot gruppierte Statistiken
qualitativ	diskret. Mosaikplot diskret. Kontigenztafel	Mosaikplot Kontigenztafel

Explorative Datenanalyse

2 quantitative Variablen

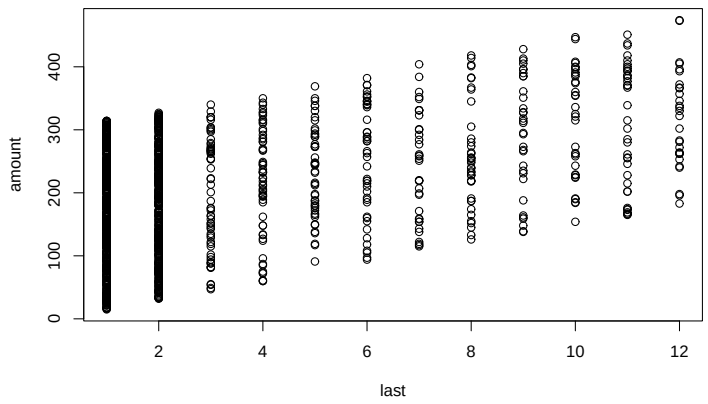
Numerische Beschreibung: Korrelationskoeffizient.

Grafische Beschreibung: Streudiagramm.

Beispiel: `amount` und `last` (aus `BBBClub`)

Korrelation: $r = 0.452$

Explorative Datenanalyse



Explorative Datenanalyse

1 abhängige quantitative und 1 erklärende qualitative Variable

Numerische Beschreibung: Gruppierte Statistiken.

Grafische Beschreibung: parallele Boxplots.

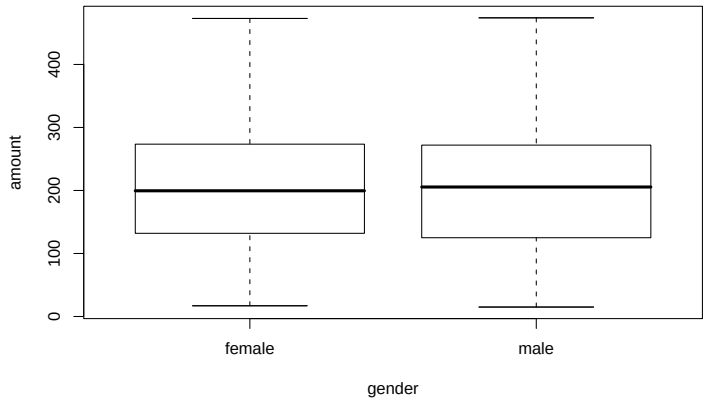
Beispiel: `amount` und `gender`

Mittelwerte: Männer 203.5, Frauen 200.2.

Fünf-Punkt-Zusammenfassung:

	Min.	Q_1	Median	Q_3	Max.
Männer	17	132	199.5	273.5	473
Frauen	15	125	205.5	272.0	474

Explorative Datenanalyse



Explorative Datenanalyse

2 qualitative Variablen

Numerische Beschreibung: Kontigenztafeln.

Grafische Beschreibung: Mosaikplot.

Beispiel: `choice` und `gender`

Geschlecht vs. Kauf	nein	ja	
Frauen	273	183	
Männer	627	217	

(absolute Häufigkeiten)

Explorative Datenanalyse

2 qualitative Variablen

Numerische Beschreibung: Kontigenztafeln, Odds Ratio.

Grafische Beschreibung: Mosaikplot.

Beispiel: `choice` und `gender`

Geschlecht vs. Kauf	nein	ja	Summe
Frauen	273	183	456
Männer	627	217	844
Summe	900	400	1,300

(absolute Häufigkeiten mit Randverteilung)

Explorative Datenanalyse

2 qualitative Variablen

Numerische Beschreibung: Kontigenztafeln, Odds Ratio.

Grafische Beschreibung: Mosaikplot.

Beispiel: `choice` und `gender`

Geschlecht vs. Kauf	nein	ja	Summe
Frauen	0.210	0.141	0.341
Männer	0.482	0.167	0.659
Summe	0.692	0.308	1

(relative Häufigkeiten mit Randverteilung)

Explorative Datenanalyse

2 qualitative Variablen

Numerische Beschreibung: Kontigenztafeln, Odds Ratio.

Grafische Beschreibung: Mosaikplot.

Beispiel: `choice` und `gender`

Geschlecht vs. Kauf	nein	ja	Summe
Frauen	0.599	0.401	1
Männer	0.743	0.257	1

(bedingte relative Häufigkeiten)

Explorative Datenanalyse

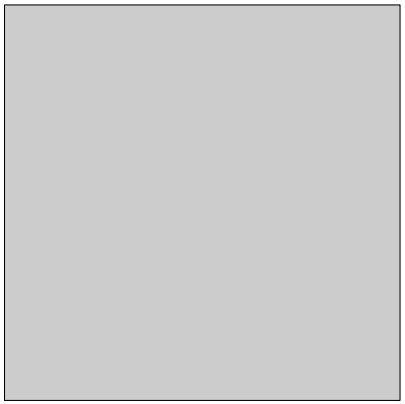
Mosaikplot

Der Mosaikplot ist eine flächenproportionale Darstellung einer Kontingenztafel, d.h. also je größer die Fläche eines Mosaiks desto größer ist der entsprechende Eintrag der Kontingenztafel.

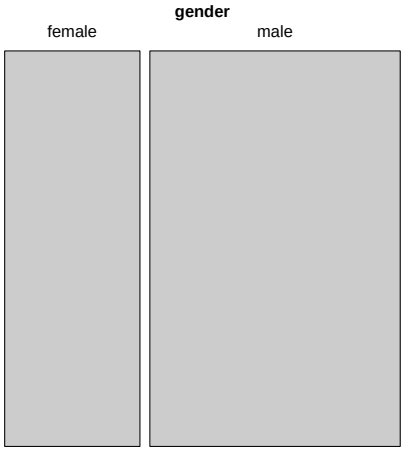
Konstruktion: eine rechteckige Fläche wird rekursiv bzgl. der relativen Häufigkeiten (bedingt auf alle vorangegangenen Ränder) geteilt.

Hier: Zunächst Split gemäß der relativen Häufigkeiten von Geschlecht. Dann Split gemäß der bedingten relativen Häufigkeiten von Kauf gegeben Geschlecht.

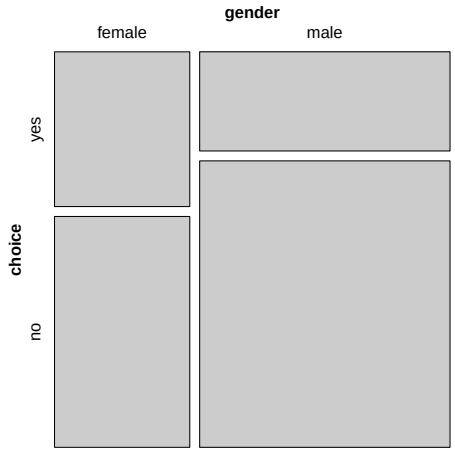
Explorative Datenanalyse



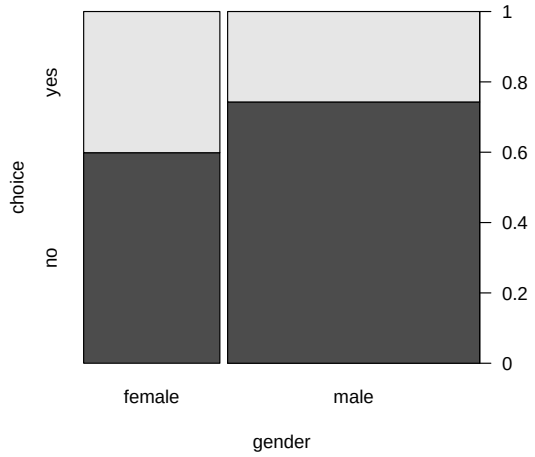
Explorative Datenanalyse



Explorative Datenanalyse



Explorative Datenanalyse



Explorative Datenanalyse

1 abhängige qualitative und 1 erklärende quantitative Variable

Numerische Beschreibung: diskretisierte Kontigenztafeln.

Grafische Beschreibung: diskretisierter Mosaikplot.

Idee: Transformiere die erklärende quantitative Variable in eine qualitative Variable durch Intervallbildung. *Dies ist dieselbe Idee wie beim Histogramm.*

Wie sollen die Intervalle gewählt werden? Beispielsweise Fünf-Punkt-Zusammenfassung als Intervallgrenzen verwenden.

Beispiel: `choice` und `amount`

Explorative Datenanalyse

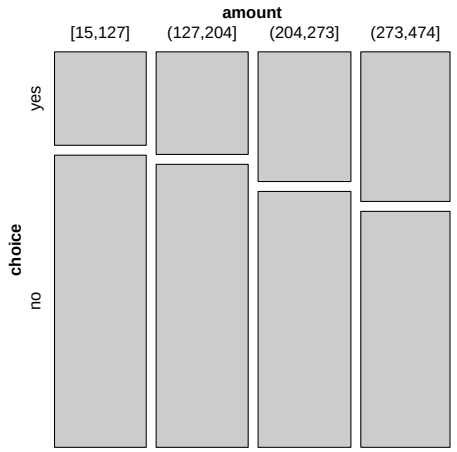
Fünf-Punkt-Zusammenfassung von `amount`:

	Min.	Q_1	Median	Q_3	Max.
Ausgaben	15	127	204	273	474

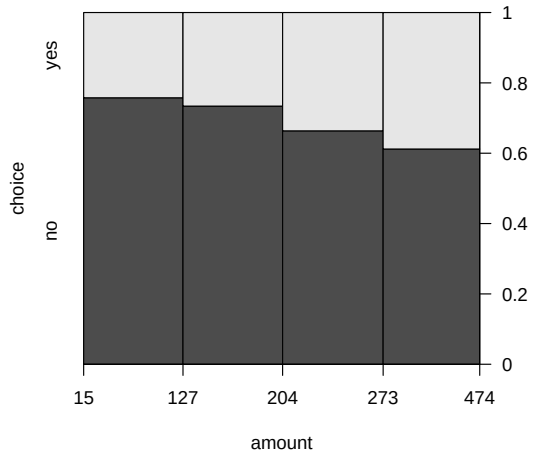
Diskretisierte Kontigenztafel:

K. vs. A.	[15, 127]	(127, 204]	(204, 273]	(273, 474]	Summe
nein	247	240	219	194	900
ja	79	87	111	123	400
Summe	326	327	330	317	1,300

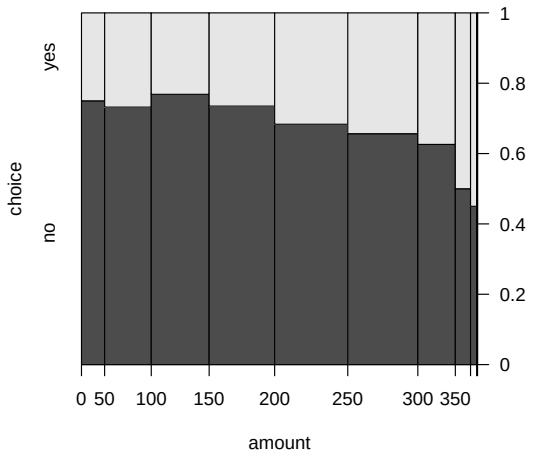
Explorative Datenanalyse



Explorative Datenanalyse



Explorative Datenanalyse



Tutorium

Bivariate explorative Datenanalyse in R ($EDA2.pdf$)

Lineare Regression

In der linearen Regression wird der Zusammenhang einer abhängige (quantitativen) Zufallsvariablen Y und k Erklärungsvariablen $X_1, \ldots, X_k$ durch eine lineare Funktion modelliert. Als erste Erklärungsvariable wird üblicherweise eine Konstante $X_1 = 1$ verwendet.

Eine Stichprobe vom Umfang n, an die ein solches Modell angepaßt werden soll, wird üblicherweise so notiert: die Beobachtungen der abhängigen Variablen y_i $(i = 1, \ldots, n)$ und eines Regressorvektors $x_i = (1, x_{i2}, \ldots, x_{ik})^\top$ $(i = 1, \ldots, n)$.

Lineare Regression

Dabei geht man davon aus, daß die Beobachtungen in Wahrheit folgender Modellgleichung genügen:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 \cdot x_{i2} + \ldots + \beta_k \cdot x_{ik} + \varepsilon_i \\ &= x_i^\top \beta + \varepsilon_i \end{aligned}$$

daß als y_i eine lineare Funktion der Regressoren x_i ist plus einen zufälligen Fehler ε_i.

Lineare Regression

In Matrixschreibweise läßt sich dies noch etwas kompakter als

$$y \quad = \quad X\beta + \varepsilon$$

notieren.

Ziel ist es nun aus den empirischen Daten y_i und x_i $(i = 1, \ldots, n)$ die Regressionskoeffizienten β möglichst gut zu schätzen, um also so das wahre Modell aus den Daten zu 'erlernen'.

Annahmen:

Dies kann aber nur (gut) funktionieren, wenn die Daten bestimmte Annahmen erfüllen. Diese Annahmen betreffen insbesonders den Modellfehler ε aber auch die Regressoren X_j.

Lineare Regression

(A1) Das Modell hat keinen systematischen Fehler.

(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß.

(A3) Die Komponenten des Fehlerterms sind nicht korreliert.

(A4) Die Regressoren sind exogen und fest vorgegeben.

(A5) Es gibt keine lin. Abhängigkeiten zwischen den Regressoren.

(A6) Der Modellfehler sei normalverteilt.

Lineare Regression

(A1) Das Modell hat keinen systematischen Fehler.
$$E(\varepsilon_i) \; = \; 0$$
(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß.
$$V(\varepsilon_i) \; = \; \sigma^2$$
(A3) Die Komponenten des Fehlerterms sind nicht korreliert.
$$COV(\varepsilon_i, \varepsilon_j) \; = \; 0$$
(A4) Die Regressoren sind exogen und fest vorgegeben.
$$COV(X_{ij}, \varepsilon_i) \; = \; 0$$
(A5) Es gibt keine lin. Abhängigkeiten zwischen den Regressoren.
$$rank(X) \; = \; k$$
(A6) Der Modellfehler sei normalverteilt.
$$\varepsilon_i \; \sim \; \mathcal{N}(0, \sigma^2)$$

Lineare Regression

Koeffizientenschätzung:

Üblicherweise werden die Koeffizienten des linearen Regressionsmodells mit der Methode der Kleinsten Quadrate (KQ) geschätzt. Genauer gesagt heißt das, daß man versucht ein Modell

$$\hat{y}_i \quad = \quad x_i^\top \hat{\beta}$$

zu finden, so daß die Fehlerquadratsumme RSS = Summe der quadrierten Residuen (oder Prognosefehler)

$$\hat{\varepsilon}_i \quad = \quad y_i - \hat{y}_i$$

Lineare Regression

minimiert wird:

$$RSS \;=\; \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \;\rightarrow\; \min$$

Der Schätzer $\hat{\beta}$, der dieses Minimierungsproblem löst, heißt KQ-Schätzer.

Lineare Regression

Man kann zeigen, daß dieser KQ-Schätzer die sogenannten Normalgleichungen

$$(X^\top X)\,\hat{\beta} \quad = \quad X^\top y$$

erfüllen muß. Wegen Annahme (A5) kann man deshalb auch schreiben

$$\hat{\beta} \quad = \quad (X^\top X)^{-1} X^\top y$$

Guass-Markov-Theorem: Unter den Annahmen (A1)–(A3) ist dies der beste lineare unverzerrte Schätzer (BLUE) für den wahren Parameter β.

Lineare Regression

Prognose:

Wenn man den Schätzer $\hat{\beta}$ für die Regressionskoeffizienten berechnet hat, ist es sehr leicht Prognosen $\hat{y}_{n+1}$ für neue Beobachtungen der Regressoren x_{n+1} zu berechnen. Man setzt alle Werte einfach in die Modellgleichung ein:

$$\begin{aligned} \hat{y}_{n+1} &= \hat{\beta}_1 + \hat{\beta}_2 \cdot x_{n+1,2} + \ldots + \hat{\beta}_k \cdot x_{n+1,k} \\ &= x_{n+1}^\top \hat{\beta} \end{aligned}$$

Lineare Regression

Inferenz:

Um zu überprüfen, ob die verwendeten Regressoren X_j überhaupt einen Einfluß auf Y haben, stehen vor allem zwei Tests zur Verfügung: der sogenannte t-Test und der F-Test.

t-Test:

Der t-Test testet für **einen** Koeffizienten die Hypothese H_0

$$H_0: \quad \beta_j \;=\; 0$$

gegen die Alternative H_1

$$H_1: \quad \beta_j \;\neq\; 0$$

Lineare Regression

Die t-Statistik ist

$$t \quad = \quad \frac{\hat{\beta}_j}{\widehat{SD(\hat{\beta}_j)}}$$

also der standardisierte Quotient des Schätzers $\hat{\beta}_j$ und seiner geschätzten Standardabweichung.

Die t-Statistik ist unter (A1)–(A6) exakt t-verteilt, ohne (A6) gilt dies approximativ (ungefähr standardnormalverteilt). Dadurch können die zugehörigen p-Werte berechnet werden.

Lineare Regression

Hieraus kann auch ein Konfidenzintervall für β_j berechnet werden. Per Faustregel ist dies

$$\hat{\beta}_j \ \pm \ 2 \cdot \widehat{SD(\hat{\beta}_j)}.$$

Lineare Regression

F-Test:

Der F-Test vergleicht genestete Modelle: das Ausgangsmodell mit k Parametern $\beta_1, \ldots, \beta_k$ und ein vereinfachtes Modell mit nur $k-q$ Parametern $\beta_1, \ldots, \beta_{k-q}$, in dem also die die übrigen q Koeffizienten 0 sind: $\beta_{k-q+1} = \ldots = \beta_k = 0$.

Die Frage ist, ob sich das Modell signifikant verschlechtert, wenn man die letzten q Parameter wegläßt (gleich 0 setzt).

Lineare Regression

Der F-Test testet für q Koeffizienten die Hypothese H_0

$$H_0: \quad \beta_{k-q+1} = \ldots = \beta_k = 0$$

gegen die Alternative H_1

$$H_1: \quad \text{mindestens ein } \beta_j \neq 0 \quad (j = k-q+1, \ldots, k)$$

Um dies zu überprüfen wird das volle Modell $\hat{y}_i$ mit k Parametern $\hat{\beta}_1, \ldots, \hat{\beta}_k$ und zugehöriger Fehlerquadratsumme RSS_1 berechnet. Dann wird das vereinfachte Modell $\tilde{y}_i$ mit k Parametern $\tilde{\beta}_1, \ldots, \tilde{\beta}_{k-q}$ und zugehöriger Fehlerquadratsumme RSS_2 berechnet.

Lineare Regression

Um zu bestimmen, ob RSS_1 signifikant kleiner ist als RSS_2 betrachtet man dann die F-Statistik

$$F = \frac{(RSS_2 - RSS_1)/q}{RSS_1/(n-k)}$$

Falls $q = 1$ ist der F-Test äquivalent zum entsprechenden t-Test, d.h. beide liefern denselben p-Wert. In diesem Fall ist $F = t^2$.

Lineare Regression

Bestimmtheitsmaß:

Um zu berechnen, welcher Anteil der Streuung von y_i durch die Regressoren x_i erklärt werden kann, läßt sich das Bestimmtheitsmaß R^2 berechnen.

Tutorium

Lineare Modelle in R: Klassische lineare Regression (*LiMo1.pdf*)

Übung

Aufgabe 2:

Betrachten Sie die die ersten 9 Variablen der Guest Survey Austria Daten (`GSA`). D.h. also *ohne* die `SA`nn`.`xxx und `M`nn`.`xxx Variablen.

Führen Sie bivariate explorative Analysen durch mit den Ausgaben `expenditure` als abhängiger Variablen und jeweils jeder der verbleibenden 8 als Erklärungsvariablen. Berücksichtigen Sie dabei wieder geeignete Transformationen der Daten.

Übung

Aufgabe 3: `Machinco`

Die Firma Machinco stellt Hoch-Technologie-Komponenten her und hat momentan 17 Kunden. Der Datensatz *Machinco.rda* gibt für jeden dieser Kunden die Anzahl der Angestellten (`employees`) und das Kaufvolumen (in USD 1000) bei Machinco (`sales`) an. Die Anzahl der Angestellten kann als grober Prädiktor für das Verkaufspotential verwendet werden.

Übung

Passen sie ein lineares Regressionsmodell an, das das Kaufvolumen durch die Anzahl der Angestellten erklärt.

- Wie lautet die geschätze Regressionsgerade?
- Visualisieren Sie die Daten und das angepaßte Modell.
- Hat die Anzahl der Angestellten einen signifikanten Einfluß auf das Kaufvolumen?
- Geben Sie ein Konfidenzintervall für den zugehörigen Regressionskoeffizienten an.
- Welches Kaufvolumen prognostizieren Sie bei einem Kunden mit 1500 Angestellten?

Schätzung im linearen Regressionsmodell

Für das lineare Regressionsmodell

$$y \quad = \quad X\beta + \varepsilon$$

ist der KQ-Schätzer und die zugehörigen Residuen gegeben durch

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top y \\ \hat{\varepsilon} &= y - \hat{y} = y - X\hat{\beta} \end{aligned}$$

Schätzung im linearen Regressionsmodell

(A1) Das Modell hat keinen systematischen Fehler.

$$E(\varepsilon_i) = 0$$

(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß.

$$V(\varepsilon_i) = \sigma^2$$

(A3) Die Komponenten des Fehlerterms sind nicht korreliert.

$$COV(\varepsilon_i, \varepsilon_j) = 0$$

(A4) Die Regressoren sind exogen und fest vorgegeben.

$$COV(X_{ij}, \varepsilon_i) = 0$$

(A5) Es gibt keine lin. Abhängigkeiten zwischen den Regressoren.

$$rank(X) = k$$

(A6) Der Modellfehler sei normalverteilt.

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Schätzung im linearen Regressionsmodell

Dabei werden in dieser LV die Annahmen über die Regressoren (A4) und (A5) **immer** vorausgesetzt:

- (A4) ist bei Querschnittsdaten in aller Regel erfüllt,
- (A5) wird zur Identifizierbarkeit des Modells benötigt (und in Einheit 4 nochmal thematisiert).

Die Annahmen, die den Modellfehler ε betreffen, werden im folgenden nochmal genauer betrachtet.

Schätzung im linearen Regressionsmodell

Als minimale Voraussetzung für ein sinnvolles Modell wird die Annahme (A1) benötigt, so daß die Fehlerterme ε_i den Mittelwert 0 haben und sich die Modellgleichung

$$y_i \quad = \quad x_i^\top \beta + \varepsilon_i$$

völlig äquivalent umformen läßt zu einer Gleichung für den Erwartungswert μ_i von y_i:

$$E(y_i) \quad = \quad \mu_i \quad = \quad x_i^\top \beta$$

Schätzung im linearen Regressionsmodell

Das heißt also, daß die Regressoren x_i und zugehörigen Regressionskoeffizienten die den Mittelwert der y_i spezifizieren.

Diese Annahme reicht bereits aus, um die Parameter β sinnvoll durch die Methode der kleinsten Quadrate (KQ) zu schätzen.

KQ-Schätzung

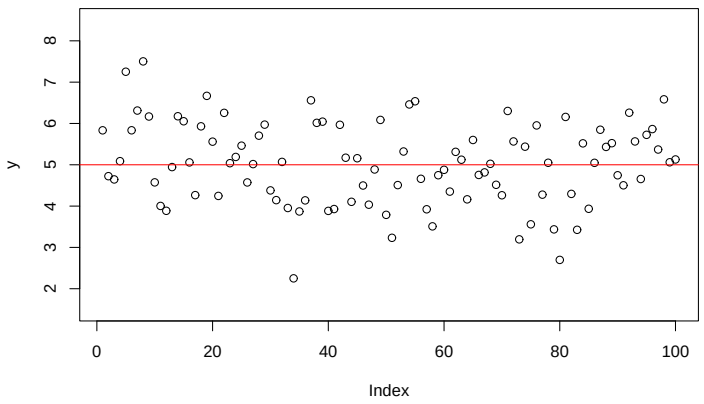
In einem sehr einfachen Modell soll dieser Schätzer einmal analytisch berechnet werden. Wenn wir annehmen, daß wir nur ein einzigen konstanten Regressor $x_i = 1$ haben, dann vereinfacht sich das Modell

$$E(y_i) \quad = \quad 1 \cdot \beta_1 \quad = \quad \mu$$

unabhängig von i.

Anschaulich gesprochen heißt ein einziger konstanter Regressor also: alle Beobachtungen y_i haben dasselbe Mittel $\mu = \beta_1$, um das sie zufällig (gemäß der Fehlergröße ε) schwanken.

KQ-Schätzung



KQ-Schätzung

Gesucht wird also ein Schätzer für den Mittelwert $\hat{\mu} = \hat{\beta}$, der damit gleichzeitig die Prognose $\hat{y}$ ist.

Das Prinzip der kleinsten Quadraten ermittelt diesen Schätzer durch Minimierung der Fehlerquadratsumme RSS in Abhängigkeit von μ bzw. β.

$$RSS(\mu) \quad = \quad \sum_{i=1}^{n}(y_i - \mu)^2 \;\rightarrow\; \min$$

Diese Minimierung wird analytisch durchgeführt, d.h. eine notwendige Bedingung ist, daß die Ableitung von RSS nach μ gleich 0 ist.

KQ-Schätzung

$$\begin{aligned}
\frac{\partial RSS(\mu)}{\partial \mu} &\stackrel{!}{=} 0 \\
\Leftrightarrow \quad \sum_{i=1}^{n} 2 \cdot (y_i - \mu) &= 0 \\
\Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \mu) &= 0 \\
\Leftrightarrow \quad \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \mu &= 0
\end{aligned}$$

KQ-Schätzung

$$\begin{aligned}
\Leftrightarrow \quad 0 &= \sum_{i=1}^{n} y_i - n \cdot \mu \\
\Leftrightarrow \quad n \cdot \mu &= \sum_{i=1}^{n} y_i \\
\Leftrightarrow \quad \mu &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\Leftrightarrow \quad \mu &= \bar{y}
\end{aligned}$$

KQ-Schätzung

Der Schätzer, der also die Fehlerquadratsumme minimiert[1], ist das arithmetische Mittel $\bar{y}$.

Dies ist der aus Statistik 1 bekannte Schätzer, der sich auch als Spezialfall der allgemeinenen Formel für k Regressoren ergibt:

[1] Die zweite Ableitung müßte streng genommen auch noch überprüft werden.

KQ-Schätzung

$$\begin{aligned}
\hat{\beta} &= (X^\top X)^{-1} X^\top y \\
&= ((1 \cdots 1)(1 \cdots 1)^\top)^{-1} (1 \cdots 1) y \\
&= \left(\sum_{i=1}^{n} 1 \right)^{-1} \sum_{i=1}^{n} y_i \\
&= n^{-1} \sum_{i=1}^{n} y_i \\
&= \bar{y}
\end{aligned}$$

Lineare Regression unter Normalverteilung

Zusätzlich zur korrekten Spezifikation des Mittelwertes (A1) unterstellt man üblicherweise konstante Varianzen (A2), Unabhängigkeit der Beobachtungen (A3) und ggf. auch Normalverteilung (A6).

Diese Annahmen lassen sich auch zusammenfassen als:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ u.i.v.}$$

unabhängig identisch verteilt. Völlig äquivalent läßt sich schreiben:

$$y_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2) \text{ u.i.v.}$$

d.h. zusätzlich zu dem Mittelwert $\mu_i = x_i^\top \beta$ (A1) ist nun die gesamte Form der Verteilung von y_i festgelegt (A2, A3, A6).

Lineare Regression unter Normalverteilung

Diese zusätzliche Information kann zur Schätzung des Parametervektors β ausgenutzt werden.

Das verwendete Schätzverfahren heißt Maximum Likelihood (ML) Schätzung.

Bevor die ML-Schätzung für normalverteilte Beobachtungen durchgeführt wird, wird sie bei binären Beobachtungen illustriert.

ML-Schätzung

Exkurs: Sei y_i nun eine binäre Variable, d.h. sie hat nur zwei Ausprägungen. Mit "Erfolgswahrscheinlichkeit" p ist $y_i = 1$ ("Erfolg") und der entsprechenden Gegenwahrscheinlichkeit $1 - p$ ist $y_i = 0$ ("Misserfolg").

$$\begin{aligned} P(Y_i = y_i) &= p^{y_i} \cdot (1-p)^{1-y_i} \\ &= \begin{cases} p & \text{falls } y_i = 1 \\ 1-p & \text{falls } y_i = 0 \end{cases} \end{aligned}$$

ML-Schätzung

Bei n unabhängigen Beobachtungen Y_i ist die gemeinsame Wahrscheinlichkeit das Produkt der Einzelwahrscheinlichkeiten, d.h. bei gegebener Erfolgswahrscheinlichkeit p:

$$\begin{aligned} P(Y = y \mid p) &= \prod_{i=1}^{n} P(Y_i = y_i) \\ &= p^{\sum_{i=1}^n y_i} \cdot (1-p)^{\sum_{i=1}^n (1-y_i)} \end{aligned}$$

ML-Schätzung

Bei der Parameterschätzung soll aber nicht bei gegebenen Parameter p eine Wahrscheinlichkeit für die Beobachtungen ausgerechnet werden. Vielmehr soll bei gegebenen Beobachtungen ein "guter" Schätzer für p gefunden werden.

Faßt man diese gemeinsame Wahrscheinlichkeit als Funktion der Parameter auf, nennt man sie **Likelihood**.

$$L(p \mid y_1, \ldots, y_n) = p^{\sum_{i=1}^n y_i} \cdot (1-p)^{n - \sum_{i=1}^n y_i} \rightarrow \max$$

Diese soll in Abhängigkeit des Parameters maximiert werden. Die Maximalstelle $\hat{p}$ ist der ML-Schätzer.

ML-Schätzung

Es ist sowohl analytisch als auch numerisch einfacher die logarithmierte Likelihood $\log L(p)$ zu maximieren. Da dies eine monotone Transformation ist, ist die Maximalstelle $\hat{p}$ bei $L(\cdot)$ und $\log L(\cdot)$ dieselbe.

$$\log L(p \mid y_1, \ldots, y_n) = \sum_{i=1}^{n} y_i \cdot \log p + \left(n - \sum_{i=1}^{n} y_i \right) \cdot \log(1-p)$$

Die notwendige Bedingung für ein Maximum ist wiederum, daß die Ableitung nach p gleich 0 ist.

ML-Schätzung

$$\begin{aligned}
& & \frac{\partial \log L(p)}{\partial p} &\stackrel{!}{=} 0 \\
\Leftrightarrow & & \sum_{i=1}^{n} y_i \cdot \frac{1}{p} - (n - \sum_{i=1}^{n} y_i) \cdot \frac{1}{1-p} &= 0 \\
\Leftrightarrow & & \sum_{i=1}^{n} y_i \cdot (1-p) - (n - \sum_{i=1}^{n} y_i) \cdot p &= 0 \\
\Leftrightarrow & & \sum_{i=1}^{n} y_i \cdot - \sum_{i=1}^{n} y_i \cdot p - n \cdot p + \sum_{i=1}^{n} y_i \cdot p &= 0
\end{aligned}$$

ML-Schätzung

$$\begin{aligned}
\Leftrightarrow \quad n \cdot p &= \sum_{i=1}^{n} y_i \\
\Leftrightarrow \quad p &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
&= \bar{y}
\end{aligned}$$

Damit ist also der ML-Schätzer für die Erfolgswahrscheinlichkeit wieder das arithmetische Mittel, d.h. der empirische Anteil der Erfolge in der Stichprobe.

ML-Schätzung

Bei diskreten (qualitativen) Beobachtungen ist wie in obigem Beispiel auch die Dichte diskret und gibt die Wahrscheinlichkeiten $P(Y_i = y_i)$ an.

Bei stetigen (quantitativen) Beobachtungen wie beispielsweise normalverteilten Beobachtungen ist die Dichte stetig, kann aber genauso zur Berechnung der Likelihood verwendet werden.

Hat also Y_i die Dichtefunktion $f(y_i \mid \beta)$ mit Parameter β, ist die (log-)Likelihood gegeben als:

ML-Schätzung

$$\begin{aligned}
L(\beta \mid y_1, \ldots, y_n) &= \prod_{i=1}^{n} f(y_i \mid \beta) \\
\log L(\beta \mid y_1, \ldots, y_n) &= \sum_{i=1}^{n} \log f(y_i \mid \beta)
\end{aligned}$$

ML-Schätzung

Zurück zum Regressionsbeispiel. Die Beobachtungen sind

$$y_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2)$$

und bei nur einem konstanten Regressor vereinfachte sich das zu

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

wobei $\mu = \beta_1$.

ML-Schätzung

Die Dichte der Normalverteilung ist:

$$f(y \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{(y-\mu)^2}{\sigma^2}\right).$$

Damit ist die log-Likelihood:

$$\begin{aligned} \log L(\mu, \sigma^2 \,|\, y_1, \ldots, y_n) &= \sum_{i=1}^{n} \log f(y_i \,|\, \mu, \sigma^2) \\ &= \sum_{i=1}^{n} \frac{(y-\mu)^2}{\sigma^2} - \frac{n}{2} \cdot \log\left(2\pi\sigma^2\right) \end{aligned}$$

ML-Schätzung

Maximierung von $\log L$ bzgl. μ:
Berechne die Ableitung von $\log L$ nach μ

$$\frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} 2 \cdot \frac{y_i - \mu}{\sigma^2}$$

und setze diese $= 0$. Dann löse nach μ.

ML-Schätzung

$$\begin{aligned} \frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} &\stackrel{!}{=} 0 \\ \Leftrightarrow \quad \sum_{i=1}^{n} 2 \cdot \frac{y_i - \mu}{\sigma^2} &= 0 \\ \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \mu) &= 0 \\ \Leftrightarrow \quad \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \mu &= 0 \end{aligned}$$

ML-Schätzung

$$
\begin{aligned}
\Leftrightarrow \quad 0 &= \sum_{i=1}^{n} y_i - n \cdot \mu \\
\Leftrightarrow \quad n \cdot \mu &= \sum_{i=1}^{n} y_i \\
\Leftrightarrow \quad \mu &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\Leftrightarrow \quad \mu &= \bar{y}
\end{aligned}
$$

ML-Schätzung

Und damit ist auch der ML-Schätzer $\hat{\mu} = \bar{y}$ das arithmetische Mittel der Beobachtungen.

Generell (und nicht nur für den Spezialfall $x_i = 1$) gilt im linearen Regressionsmodell mit normalverteilten Fehlern, daß KQ-Schätzer und ML-Schätzer identisch sind

$$\hat{\beta} \quad = \quad (X^\top X)^{-1} X^\top y$$

Im Allgemeinen gilt das **nicht**. Insbesonders in allen übrigen Modellen in diesem Teil der LV unterscheiden sich KQ-Schätzer und ML-Schätzer. Dann ist der ML-Schätzer üblicherweise zu bevorzugen.

Schätzung im linearen Regressionsmodell

Unter (A1): KQ-Schätzung ist sinnvoll (konsistent). Falls (A1) verletzt ist, ist das Modell nicht sinnvoll. Beispiele: vergessene Regressoren, nicht-lineare Beziehung zwischen x_i und y_i, Strukturveränderungen uvm.

Falls zusätzlich (A2) und (A3): KQ-Schätzer ist optimal (BLUE). Falls (A2) verletzt ist: Transformationen, andere Varianzschätzer, u.a.

Falls zusätzlich (A6): KQ-Schätzer ist auch der ML-Schätzer.

Tutorium

Einfaches Datenmanagement in R ($Daten.pdf$)

Übung

Aufgabe 4:

Gegeben seien n unabhängige Beobachtungen $y_1, \ldots, y_n$ aus einer Poisson-Verteilung. Diese Verteilung hat die Dichtefunktion

$$f(y \,|\, \lambda) \quad = \quad \frac{\lambda^y \cdot \exp(-\lambda)}{y!}$$

mit Parameter λ. Dieser beschreibt den Erwartungswert der Variablen, es gilt: $E(y_i) = \lambda$.

- Wie lautet der KQ-Schätzer für λ?
- Wie lautet der ML-Schätzer für λ?

Übung

Aufgabe 5:

Wählen Sie aus dem Datensatz `GSA` die Touristen eines Landes (`country`) Ihrer Wahl aus.

Passen Sie für diese Teilstichprobe des Datensatzes ein lineares Regressionsmodell an, das den Logarithmus der Ausgaben (`log(expenditure)`) versucht durch den Logarithmus des Einkommens (`log(income)`) und durch das Alter (`age`) zu erklären.

Welche der beiden Regressorvariablen hat einen signifikanten Einfluß auf die Ausgaben?

Einweg-Varianzanalyse

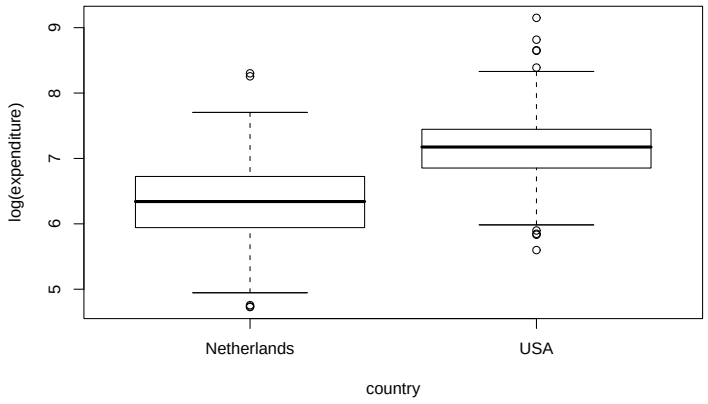
Bisher haben wir im linearen Regressionsmodell

$$\begin{aligned} y_i &= \beta_1 + \beta_2 \cdot x_{i2} + \ldots + \beta_k \cdot x_{ik} + \varepsilon_i \\ y_i &= x_i^\top \beta + \varepsilon_i \\ E(y_i) = \mu_i &= x_i^\top \beta \end{aligned}$$

als Erklärungsvariablen immer nur quant. Variablen x_{ij} verwendet. Aber natürlich können auch qual. Erklärungsvariablen einen Einfluß auf y_i haben.

In der explorativen Datenanalyse haben wir dies Problem auch schon durch parallele Boxplots visualisiert.

Einweg-Varianzanalyse



Einweg-Varianzanalyse

Intuitiv klar: Um zu prüfen, ob die Erklärungsvariable `country` einen Einfluß auf den Erwartungswert von `log(expenditure)` ($= y$) hat, vergleicht man die empirischen Mittelwerte der verschiedenen Länder: 6.331 (Niederlande) und 7.157 (USA).

Aus Statistik I bekannt: Um formal zu testen, ob sich diese Mittelwerte auch *signifikant* unterscheiden, kann ein 2-Stichproben t-Test verwendet werden. Idee der Teststatistik: Differenz der Mittelwerte ($7.157 - 6.331 = 0.826$) standardisiert durch geeignete Standardabweichung.

Eventuell bekannt: Dieser Test ist i.W. äquivalent zu einer Varianzanalyse (ANOVA) und dem zugehörgien linearen Modell.

Einweg-Varianzanalyse

Problem: Um ein lineares Regressionsmodell anzupassen, brauchen wir numerische Erklärungsvariablen x_{ij}, wir haben jedoch nur eine qualitative Variable. Die Variable `country` hat Ausprägungen (Niederlande, Niederlande, USA, Niederlande, USA, $\dots$)$^\top$.

Idee: Erzeuge zwei sogenannte Indikatorvariablen oder Dummy-Variablen, die anzeigen, ob eine bestimmte Beobachtung zu den Niederlanden gehört $(1, 1, 0, 1, 0, \dots)^\top$ bzw. zu den USA $(0, 0, 1, 0, 1, \dots)^\top$ und benutze diese in der Regression:

$$y_i \quad = \quad \beta_1 + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \varepsilon_i$$

Einweg-Varianzanalyse

In Matrixschreibweise:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ & \vdots & \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{pmatrix}$$

Einweg-Varianzanalyse

Problem: Diese Regressormatrix verletzt die Annahme (A5):

> Es gibt keine linearen Abhängigkeiten zwischen den Regressoren. Hier: $rank(X) = 3$.

Hier gibt es aber eine ganz offensichtliche lineare Abhängigkeit, die erste Spalte ist nämlich die Summe der übrigen zwei Spalten. Deshalb ist hier $rank(X) = 2$, was für praktische Zwecke bedeutet: man kann nur zwei Parameter (und eben nicht drei) schätzen.

Dies ist auch intuitiv, da es ja nur zwei Gruppen gibt, und entsprach auch unserern ursprünglichen Idee die zwei Mittelwerte zu schätzen.

Einweg-Varianzanalyse

Lösung: Es wird eine lineare Nebenbedingung an die drei Parameter

- $\beta_1 =$ Achsenabschnitt
- $\beta_2 =$ Niederlande-Effekt
- $\beta_3 =$ USA-Effekt

angelegt, um das Modell identifizierbar zu machen. Hierfür gibt es verschiedene Möglichkeiten, unter anderem:

- $\beta_1 = 0$,
- $\beta_2 = 0$,
- $\beta_2 + \beta_3 = 0$.

Restriktionen, die sich nur auf die Koeffizienten beziehen, die zur Erklärungsvariable gehören, (hier β_2, β_3) nennt man auch Kontraste.

Einweg-Varianzanalyse

Die Restriktion $\beta_1 = 0$ führt dazu, dass kein Achsenabschnitt mitgeschätzt wird. Die beiden übrigen Schätzer sind mit

$$\hat{\beta}_2 = 6.331, \quad \hat{\beta}_3 = 7.157$$

einfach die Gruppenmittel.

Wir sind aber eigentlich nicht an den Gruppenmitteln selbst interessiert, sondern vielmehr an ihrer Differenz (und ob diese signifikant ist). Deshalb ist es in der Regel günstiger den Kontrast $\beta_2 = 0$ zu verwenden. Als Koeffizientenschätzungen ergeben sich dann

$$\hat{\beta}_1 = 6.331, \quad \hat{\beta}_3 = 0.826.$$

Einweg-Varianzanalyse

Der Achsenabschnitt $\hat{\beta}_1$ schätzt also den Mittelwert in der Niederlande-Stichprobe und der Landeseffekt $\hat{\beta}_3$ die Differenz zwischen den beiden Mittelwerten.

Ein Test auf die Signifikanz von $\hat{\beta}_3$ (d.h. ob β_3 von 0 verschieden ist), ist genau der Test, der prüft, ob `country` einen Einfluß auf `log(expenditure)` hat.

Diese Kontraste heißen auch Treatment-Kontraste.

Einweg-Varianzanalyse

Die Prognosen des Modells sind übrigens davon unabhängig, welche Kontraste verwendet werden, es sind immer die entsprechenden Gruppenmittelwerte.

Den Test, der prüft, ob `country` einen Einfluß auf `log(expenditure)` hat, nennt man auch Einweg-Varianzanalyse.

Einweg-Varianzanalyse

Frage: Die Einweg-Varianzanalyse ist äquivalent zum 2-Stichproben t-Test, warum braucht man das komplizierte lineare Modell?

Antwort: Dieselben Ideen zur Einbettung in lineare Modelle funktionieren auch, wenn man mehrere qualitative Erklärungsvariablen hat (Mehrweg-Varianzanalyse) bzw. gemischte quantitative und qualitative Erklärungsvariablen (Kovarianzanalyse).

Tutorium

Lineare Modelle in R: Einweg-Varianzanalyse ($LiMo2.pdf$)

Übung

Aufgabe 6:

Wählen Sie aus dem Datensatz `GSA` die Touristen eines Landes (`country`) Ihrer Wahl aus.

Passen Sie für diese Teilstichprobe des Datensatzes ein lineares Modell an, das die Abhängigkeit des Logarithmus der Ausgaben (`log(expenditure)`) vom Zielbundesland (`province`) modelliert.

Hat `province` einen signifikanten Einfluß auf `log(expenditure)`? Falls ja, beschreiben/visualisieren sie diesen Einfluß.

Zweiweg-Varianzanalyse

Bisher haben wir in der Einweg-Varianzanalyse den Einfluß einer kategorialen Variablen mit k verschiedenen Ausprägungen auf eine numerische Variable betrachtet.

Die Idee war, daß die erklärende Variable die Gesamtstichprobe in k Teilstichproben teilt, und in jeder dieser Stichproben kann ein Mittelwert geschätzt werden. In der Übertragung ins lineare Modell wurden außerdem geeignete Kontraste zur Schätzung der Parameter gewählt.

Zweiweg-Varianzanalyse

Frage: Was passiert, wenn es nun zwei qualitative Erklärungsvariablen mit k bzw. ℓ Ausprägungen gibt?

Antwort: Dann fällt jede Beobachtung der abhängigen Variablen in eine von $k \cdot \ell$ Teilstichproben, die durch Kombination der beiden Erklärungsvariablen entstehen.

Konsequenz: In so einer Situation können also bis zu $k \cdot \ell$ Parameter geschätzt werden.

Zweiweg-Varianzanalyse

Künstliches Beispiel: Modelliere den Erwartungswert von `y` durch zwei Faktoren: `a`, der die beiden Ausprägungen 'schwarz' und 'rot' hat, sowie `b`, der die beiden Ausprägungen 'ja' und 'nein' hat.

Es gibt also vier mögliche Kombinationen der Erklärungsvariablen mit den zugehörigen Erwartungswerten von `y`.

a vs. b	**ja**	**nein**
rot	$\mu_{\text{rot,ja}}$	$\mu_{\text{rot,nein}}$
schwarz	$\mu_{\text{schwarz,ja}}$	$\mu_{\text{schwarz,nein}}$

Zweiweg-Varianzanalyse

Frage: Wie hängt der Erwartungswert μ von `y` von den Variablen `a` und `b` ab?

Trivial: Wenn alle Erwartungswerte gleich sind

$$\mu_{\text{rot,ja}} = \cdots = \mu_{\text{schwarz,nein}}$$

dann hängt `y` weder von `a` noch `b` ab. Der für alle gleiche Erwartungswert kann dann geschätzt werden durch den Gesamtmittelwert.

In R: `lm(y ~ 1)`

Zweiweg-Varianzanalyse

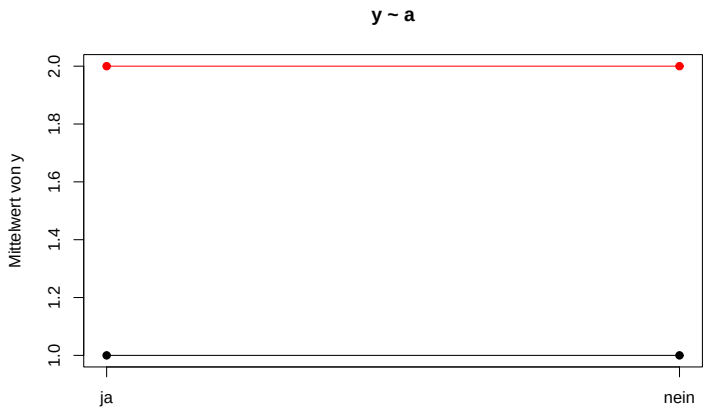
Außerdem klar: Wenn `a` einen Einfluß auf `y` hat, aber nicht `b`, dann unterscheiden sich die Zeilen in der Matrix der Mittelwerte aber nicht die Spalten.

In R: `lm(y ~ a)`

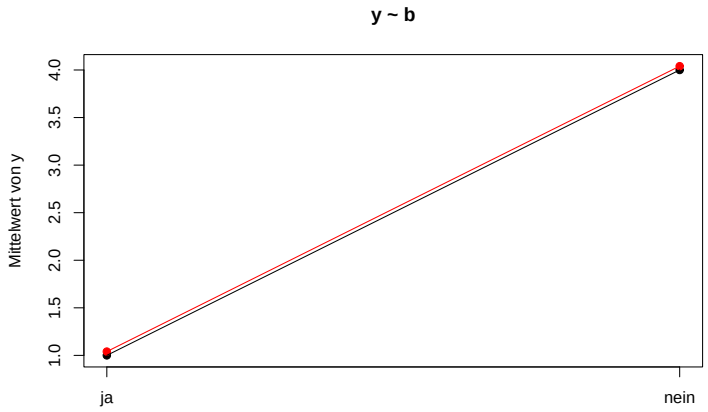
Analog: Wenn `b` einen Einfluß auf `y` hat, aber nicht `a`, dann unterscheiden sind die Spalten in der Matrix der Mittelwerte aber nicht die Zeilen.

In R: `lm(y ~ b)`

Zweiweg-Varianzanalyse



Zweiweg-Varianzanalyse



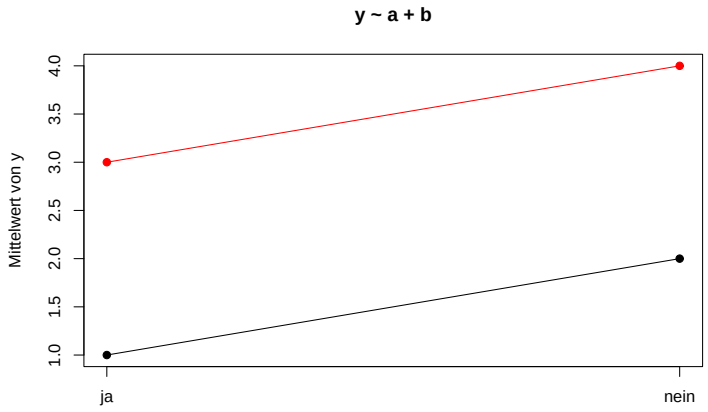
Zweiweg-Varianzanalyse

Man kann bei qualitativen Erklärungsvariablen, genau wie bei quantitativen Erklärungsvariablen, auch den Effekt mehrerer Variablen gleichzeitig schätzen.

In R: `lm(y ~ a + b)`

Hier gibt es also einen ‘rot’-Effekt und einen ‘nein’-Effekt. Beide Effekte sind additiv. Das heißt der ‘rot’-Effekt ist unabhängig davon, ob die Variable b den Wert ‘ja’ oder ‘nein’ annimmt – und umgekehrt ist der ‘nein’-Effekt unabhängig davon, ob die Variable a den Wert ‘schwarz’ oder ‘rot’ annimt.

Zweiweg-Varianzanalyse



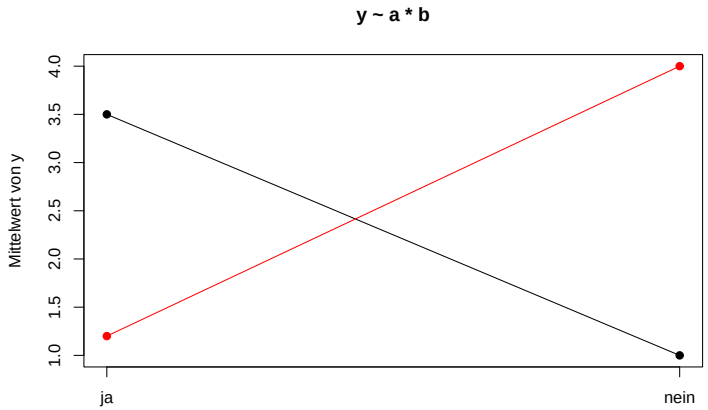
Zweiweg-Varianzanalyse

Wenn aber der 'rot'-Effekt und der 'nein'-Effekt nicht unabhängig sind, dann liegt eine Interaktion zwischen den beiden Variablen a und b vor. Das heißt also, dass der 'nein'-Effekt davon abhängt, ob die Variable a den Wert 'schwarz' oder 'rot' annimmt. Und umgekehrt hängt der 'rot'-Effekt davon ab, ob die Variable b den Wert 'ja' oder 'nein' annimmt.

In R: `lm(y ~ a + b + a:b)`

Die Effekte a und b nennt man Haupteffekte und a:b die Interaktion. Eine verkürzte Schreibweise für beides zusammen ist: `lm(y ~ a * b)`

Zweiweg-Varianzanalyse



Zweiweg-Varianzanalyse

In dem Modell mit Interaktion

`y ~ a + b + a:b`

wird die maximale Anzahl von Parameter, also $2 \cdot 2 = 4$, geschätzt.

Frage: Wie wird dabei der Effekt der Interaktion geschätzt?

Klar: Der erste Koeffizient ist wieder der Achsenabschnitt, der also dem Mittelwert in der Gruppe 'schwarz und ja' entspricht. Im linearen Modell wird a zu einer Indikatorvariable für den 'rot'-Effekt transformiert und b zu einer Indikatorvariable für den 'nein'-Effekt.

Zweiweg-Varianzanalyse

Lösung: Falls beide Effekte gleichzeitig eintreten, also 'rot und nein', dann wird noch ein zusätzlicher Effekt geschätzt. Dies ist der Interaktionseffekt, die zugehörige Indikatorvariable ist einfach das Produkt der 'rot'- bzw. 'nein'-Indikatorvariable.

Zu beachten: Ein Interaktionseffekt macht nur dann Sinn, wenn auch die beiden zugehörigen Haupteffekte im Modell vertreten sind. Ein Modell wie

`y ~ a + a:b`

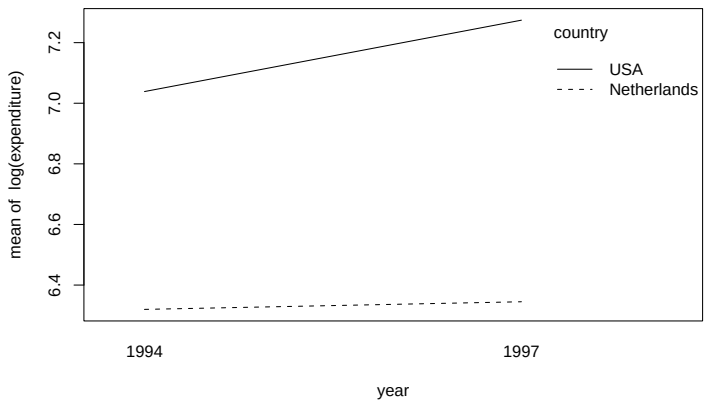
anzupassen ist deshalb zwar technisch möglich aber inhaltlich (fast) nie sinnvoll.

Zweiweg-Varianzanalyse

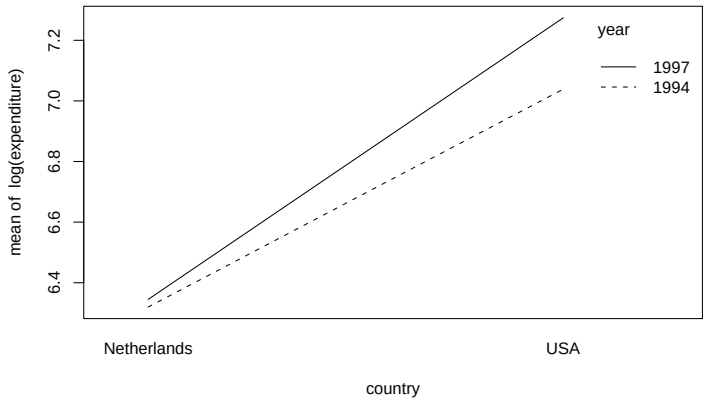
Beispiel: Abhängigkeit der log-Ausgaben `log(expenditure)` von `country` (wie bei der Einweg-Varianzanalyse) und zusätzlich vom Jahr `year`. Letzteres ist hier als Faktor kodiert, weil die Daten nur in den zwei Jahren 1994 und 1997 erhoben wurden.

In den folgenden Interaktionsplots ist der empirische Mittelwert der log-Ausgaben in den vier Gruppen gegen die beiden Erklärungsvariablen abgetragen. Die Plots sind in der Aussage äquivalent, nur die Reihenfolge der Erklärungsvariablen ist vertauscht.

Zweiweg-Varianzanalyse



Zweiweg-Varianzanalyse



Kovarianzanalyse

Völlig analog zur Mehrweg-Varianzanalyse, in der man den Einfluß von mehreren qualitativen Variablen auf eine quantitative betrachtet, kann man auch Modelle betrachten, in denen es sowohl quantitative als auch qualitative Erklärungsvariablen gibt. Diesen Fall bezeichnet man auch als Kovarianzanalyse.

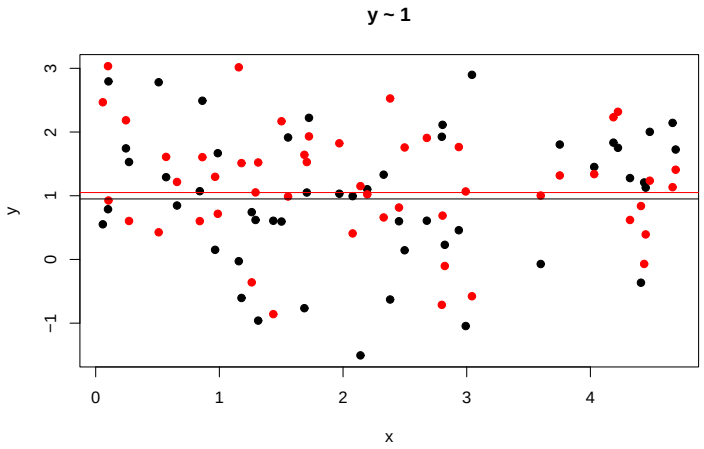
Beispiel: Es werden wieder die künstlichen Daten `y` (quantitativ) und `a` (qualitativ: schwarz, rot) betrachtet mit einer zusätzlichen quantitativen Erklärungsvariablen `x`.

Kovarianzanalyse

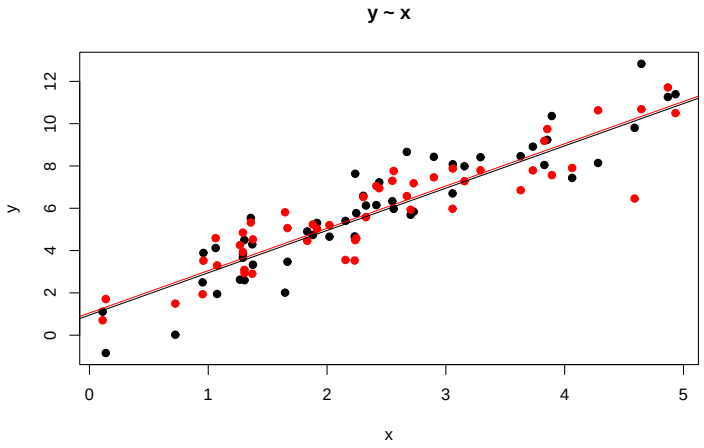
In diesem Fall ist das triviale Modell sowie die Modelle, die nur einen der Haupteffekte beinhalten bereits bekannt:

- `y ~ 1` triviales Modell (1 Mittelwert)
- `y ~ a` Einweg-Varianzanalyse (2 Mittelwerte)
- `y ~ x` Lineare Regression (1 Regressionsgerade)

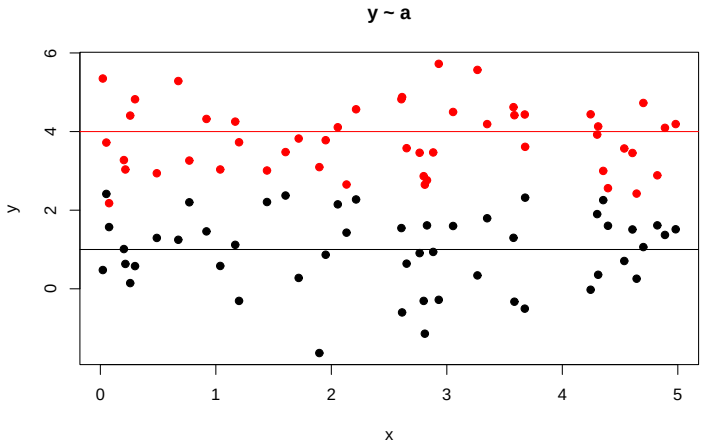
Kovarianzanalyse



Kovarianzanalyse



Kovarianzanalyse



Kovarianzanalyse

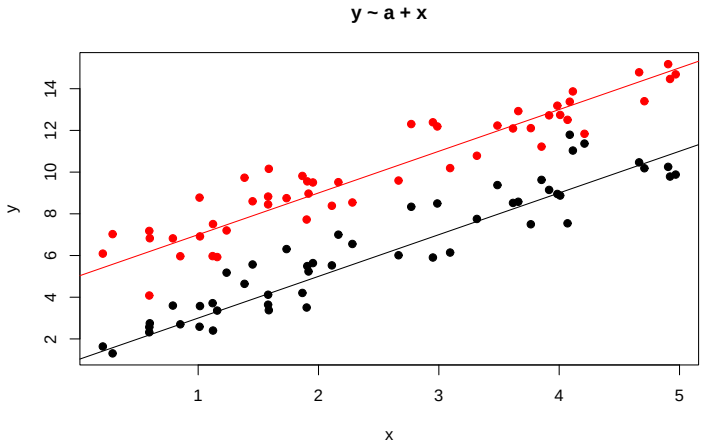
Wenn beide Erklärungsvariablen im Modell vertreten sind:

`y ~ a + x` Haupteffektmodell (2 parallele Regressionsgeraden)
Die Geraden haben unterschiedliche Achsenabschnitte gemäß `a` aber dieselbe Steigung bzgl. `x`

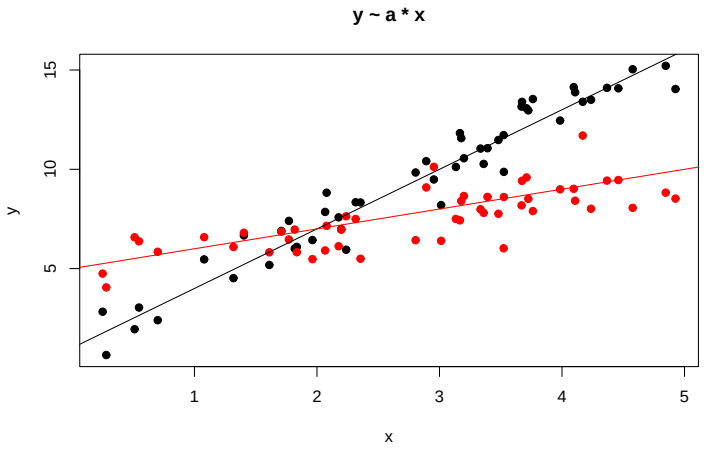
`y ~ a * x` Interaktionsmodell (2 Regressionsgeraden)
Die Geraden haben in den durch `a` bestimmten Gruppen sowohl verschiedene Achsenabschnitte als auch verschiedene Steigungen.

Das Interaktionsmodell kann wieder als `y ~ a + x + a:x` geschrieben werden. Die Interaktion `a:x` wird wieder als Produkt der Indikatorvariable von `a` und dem Regressor `x` kodiert.

Kovarianzanalyse



Kovarianzanalyse



Kovarianzanalyse

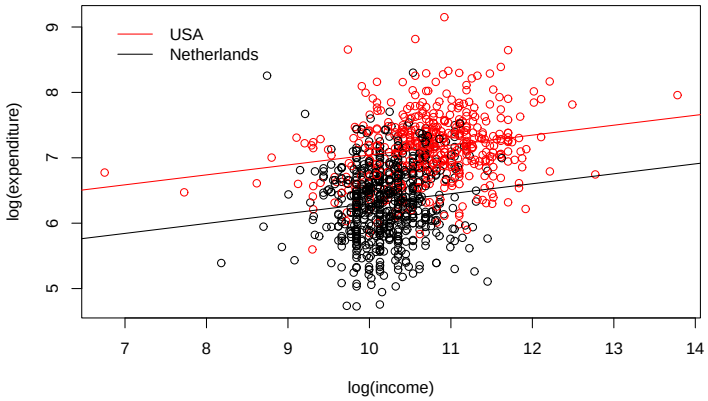
Beispiel: Betrachte die Modelle

```
log(expenditure) ~ country + log(income)
```

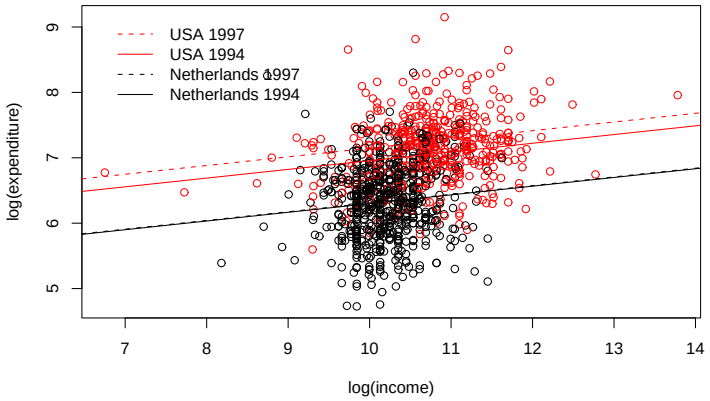
und

```
log(expenditure) ~ country * year + log(income)
```

Kovarianzanalyse



Kovarianzanalyse



Modellwahl

Um verschiedene genestete Modelle zu vergleichen, haben wir bisher vor allem den F-Test verwendet.

Erinnerung: Der F-Test vergleicht ein komplexeres Modell mit Schätzer $\hat{\beta}$ und ein vereinfachtest Teilmodell $\tilde{\beta}$. Das vereinfachte Modell hat immer ein höhere Fehlerquadratsumme, aber der F-Test prüft, ob sie auch signifikant höher ist. Wenn der p-Wert signifikant ist, verwirft man das einfache Modell und entscheidet sich für das komplexere Modell, sonst wird das einfachere Modell beibehalten.

Modellwahl

Falls sich die beiden Modelle um genau einen Parameter unterscheiden, so ist der F-Test äquivalent zum t-Test dieses Parameters.

Problem: Wenn man viele Modelle miteinander vergleicht, fällt das Konfidenzniveau rapide ab. Wenn man alle Tests zum 5% Niveau durchführt, macht man also nur mit Wahrscheinlichkeit 5% einen Fehler, wenn man die Nullhypothese verwirft. Man liegt dabei also mit 95% Wahrscheinlichkeit richtig. Wenn man allerdings 10 unabhängige Tests durchführt liegt man nur noch mit Wahrscheinlichkeit $0.95^{10} = 0.599$ richtig (d.h. für alle 10 Tests). Die Irrtumswahrscheinlichkeit ist also auf rund 40% angestiegen.

Informationskriterien

Deshalb verwendet man als Alternative zu F-Tests zur Modellwahl auch sogenannte Informationskriterien. Die Idee dafür ist folgende: Durch Hinzunahme von Erklärungsvariablen **muss** die Fehlerquadratsumme eines Modells immer fallen bzw. log-Likelihood ansteigen. Man kann also nicht allein auf Basis der Fehlerquadratsumme/log-Likelihood ein Modell wählen, dies würde **immer** das komplexeste Modell auswählen.

Als Ausgleich wird also ein Strafterm hinzuaddiert, der mit Anzahl der Parameter ansteigt. So trifft das Informationskriterium eine Entscheidung über Trade-off zwischen Minimierung der Zielfunktion (RSS, $\log L$) und der Minimierung der Anzahl der Parameter.

Informationskriterien

Zwei typischerweise verwendete Informationskriterien sind das AIC (Akaike Information Criterion) und das BIC (Bayes Information Criterion, auch SC oder SBC).

$$\begin{aligned} \mathrm{AIC}(\hat{\beta}) &= -2 \cdot \log L(\hat{\beta}) + 2 \cdot k \\ \mathrm{BIC}(\hat{\beta}) &= -2 \cdot \log L(\hat{\beta}) + \log(n) \cdot k \end{aligned}$$

wobei $\log L(\hat{\beta})$ die log-Likelihood eines Modells mit Schätzer $\hat{\beta}$ ist, k die Anzahl der Parameter und n die Anzahl der Beobachtungen.

Beide Informationskriterien entscheiden sich jeweils für das einfachere Modell, wenn $\mathrm{IC}(\tilde{\beta}) < \mathrm{IC}(\hat{\beta})$.

Informationskriterien

Sowohl AIC als auch BIC haben eine theoretische Rechtfertigung, da sie asymptotisch konsistent sind (d.h. das richtige Modell wählen).

Heuristisch lassen sich beide Kriterien folgendermaßen vergleichen:

Das AIC wählt eher zu große Modelle, versucht also ein Modell zu finden, in dem auf jeden Fall alle relevanten Erklärungsvariablen sind (aber möglicherweise auch mehr).

Das BIC hingegen bestraft die Anzahl der Parameter stärker und versucht so ein Modell zu finden, in dem alle Erklärgunsvariablen relevant sind (aber möglicherweise einige relevante Variablen vergessen wurden).

Tutorium

Lineare Modelle in R: Zweiweg-Varianzanalyse und Kovarianzanalyse ($LiMo3.pdf$)

Übung

Aufgabe 7:

Der Datensatz `vitcap` enthält Daten über die Vitalkapazität von 24 Arbeitern in der Kadmium-Industrie. Die Variable `volume` gibt das Lungenvolumen der Arbeiter (in Liter) an, ein Maß für die Vitalkapazität. Als Erklärungsvariablen stehen das Alter `age` zur Verfügung sowie ein Faktor `exposure`, der angibt ob die Arbeiter länger als 10 Jahre schädlichen Stoffen ausgesetzt waren oder nicht.

- Visualisieren Sie die Daten geeignet.
- Bestimmen Sie ein geeignetes Modell zur Erklärung von `volume`. Welches Modell wird durch F-Tests, AIC bzw. BIC gewählt?

Schrittweise Modellselektion

Bisher haben wir bei der Modellwahl immer alle relevanten Modelle angepaßt und dann miteinander verglichen. Wenn es viele potentielle Erklärungsvariablen gibt, ist dies in aller Regel zu aufwändig. Alternativ:

Vorwärtsselektion: Man startet mit dem einfachsten plausiblen Modell. Dann werden so lange Variablen in das Modell aufgenommen, wie sich das Modell verbessert. In jedem Schritt wird die Variable gewählt, die die größte Verbesserung bewirkt.

Schrittweise Modellselektion

Rückwärtsselektion: Man startet mit dem komplexesten plausiblen Modell. Es werden so lange Variablen aus dem Modell weggelassen, wie sich das Modell nicht verschlechtert. In jedem Schritt wird die Variable weggelassen, die die größte Verbesserung bewirkt.

Dabei können in jedem Schritt die Modelle entweder mit F-Tests oder einem Informationskriterium verglichen werden.

Schrittweise Modellselektion

Wird die schrittweise Modellselektion auf Basis eines Informationskriteriums durchgeführt, so kann Vorwärts- und Rückwärtsselektion auch kombiniert werden. Im ersten Schritt wählt man dabei üblicherweise ein plausibles Modell (also nicht notwendigerweise das einfachste oder komplizierteste). Dann werden alle Modelle angepaßt, in denen entweder eine verwendete Variable weggelassen oder eine nicht verwendete Variable hinzugefügt wird. Das beste all dieser Modelle wird verwendet und der Algorithmus wiederholt.

In R: `step(lmobj)`

Methoden für lm-Objekte

`summary(lmobj)`	Zusammenfassung des angepaßten Modells, enthält insbesonders Koeffizientenschätzungen und t-Tests jedes Koeffizienten sowie F-Test gegen das triviale Modell.
`coef(lmobj)`	Extraktion der Koeffizientenschätzer.
`fitted(lmobj)`	Extraktion der angepaßten Werte, d.h. Prognosen für den verwendeten Datensatz.

Methoden für lm-Objekte

`fitted(lmobj)`	Extraktion der Residuen, d.h. Prognosefehler.
`predict(lmobj, newdata)`	Prognose auf neuen Daten.
`anova(lmobj1, lmobj2)`	Modellvergleich von genesteten Modellen mit Hilfe einer Varianzzerlegung und zugehörigem F-Test.

Methoden für lm-Objekte

`logLik(lmobj)`	Extraktion der maximierten Likelihood und zugehöriger Freiheitsgrade.
`AIC(lmobj)`	Berechnung des AIC (und anderen Informationskriterien).
`step(lmobj)`	Schrittweise Modellwahl basierend auf AIC.

Dieselben Methoden sind nicht nur für lineare Modelle verfügbar, sondern auch für verallgemeinerte lineare Modelle.

Tutorium

Lineare Modelle in R: Schrittweise Modellwahl ($LiMo4.pdf$)

Das verallgemeinerte lineare Modell

Bisher haben wir das lineare Regressionsmodell betrachtet

$$\begin{aligned} y_i &= x_i^\top \beta + \varepsilon_i \\ E(y_i) = \mu_i &= x_i^\top \beta \end{aligned}$$

in dem der Erwartungswert μ_i einer metrischen abhängigen Variablen y_i durch die Linearkombination $x_i^\top \beta$ verschiedener Erklärungsvariablen mit Koeffizienten β modelliert wurde.

Dabei können die Erklärungsvariablen sowohl metrische Variablen sein als auch eine geeignete numerische Kodierung von qualitativen Variablen.

Das verallgemeinerte lineare Modell

Die Methode der kleinsten Quadrate (OLS) kann bei richtiger Spezifikation des Erwartungswertes (Annahme A1) zur Schätzung verwendet werden.

Nimmt man zusätzlich an, daß die Beobachtungen unabhängig normalverteilt sind mit derselben Varianz (Annahmen A2, A3, A6), so ist die gesamte gemeinsame Verteilung von $y = (y_1, \ldots, y_n)^\top$ spezifiziert (und nicht nur der Erwartungswert). Dies ist gleichbedeutend damit, daß die gesamte Likelihood des Modells spezifiziert ist.

In diesem Fall ist die Schätzmethode Maximum Likelihood (ML) äquivalent zu OLS.

Das verallgemeinerte lineare Modell

Wiederholung: Die Verteilung einer Variablen y_i wird beschrieben durch ihre Dichtefunktion $f(y_i \,|\, \beta)$, die möglicherweise unbekannte Parameter enthält.

Die Likelihood der Parameter ist genau dieselbe Funktion, sie wird jedoch aufgefaßt als Funktion der Parameter gegeben die Beobachtung.

$$L(\beta \,|\, y_i) \quad = \quad f(y_i \,|\, \beta).$$

Hat man nicht eine einzelne Beobachtung sondern n unabhängige Beobachtungen, so ergibt sich deren Verteilung als Produkt der einzelnen Verteilungen.

Das verallgemeinerte lineare Modell

Dementsprechend ist die Likelihood dieser Stichprobe

$$L(\beta \,|\, y_1, \ldots, y_n) \quad = \quad \prod_{i=1}^{n} f(y_i \,|\, \beta).$$

Da Summen numerisch leichter zu handhaben sind als Produkte, wird die Likelihood in aller Regel logarithmiert. Das Ergebnis einer Optimierung bleibt dabei unverändert.

$$\log L(\beta \,|\, y_1, \ldots, y_n) \quad = \quad \sum_{i=1}^{n} \log f(y_i \,|\, \beta).$$

Das verallgemeinerte lineare Modell

Nun möchten wir eine ähnliche Methode nicht nur für (approximativ) normalverteilte Daten anwenden, sondern auch für binäre Daten. Diese werden in aller Regel mit Hilfe der Binomialverteilung modelliert, es ist also

$$f(y\,|\,p) \quad = \quad p^y \cdot (1-p)^{1-y}$$

wobei p die Erfolgswahrscheinlichkeit ist und y die Werte 0 (Misserfolg) oder 1 (Erfolg) annehmen kann.

Der Erwartungswert ist $E(y) = p$.

Das verallgemeinerte lineare Modell

Um nun ein Modell für den Erwartungswert $E(y_i) = \mu_i = p_i$ einer binären Variable zu formulieren, soll wieder ein linearer Prädiktor $x_i^\top \beta$ verwendet werden.

Problem: Während der lineare Prädiktor $x_i^\top \beta$ prinzipiell alle reellen Werte annehmen kann, liegt der Erwartungswert ($=$ Erfolgswahrscheinlichkeit) $\mu_i = p_i$ immer im Intervall $[0, 1]$.

Das verallgemeinerte lineare Modell

Lösung: Verwende eine Link-Funktion g, die das Intervall $[0, 1]$ auf die reellen Zahlen abbildet, so daß

$$g(\mu_i) \quad = \quad x_i^\top \beta$$

Frage: Was ist eine *geeignete* Link-Funktion?

Exkurs: Vierfeldertafeln

In der explorativen Datenanalyse wurde bereits der Zusammenhang von zwei kategorialen Merkmalen mit Hilfe ihrer Kontingenztafel und dem zugehörigen Mosaikplot untersucht

Beispiel: `choice` und `gender` (`BBBClub` Daten)

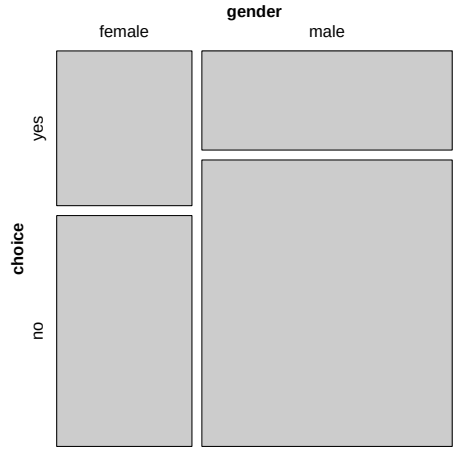
Geschlecht vs. Kauf	nein	ja
Frauen	273	183
Männer	627	217

Exkurs: Vierfeldertafeln

Um zu untersuchen, ob sich die Kaufwahrscheinlichkeit für Männer und Frauen unterscheidet, sind die bedingten relativen Häufigkeiten von Kauf gegeben Geschlecht geeigneter.

Geschlecht vs. Kauf	nein	ja	Summe
Frauen	0.599	0.401	1
Männer	0.743	0.257	1

Exkurs: Vierfeldertafeln



Exkurs: Vierfeldertafeln

Um diese vier bedingten relativen Häufigkeiten noch weiter zu komprimieren, betrachtet man anstatt der *Wahrscheinlichkeiten* die zugehörigen Chancen. Die Chance eines Ereignisses ist die Wahrscheinlichkeit für sein Eintreten geteilt durch die Gegenwahrscheinlichkeit. Im englischen heissen diese Odds.

$$\text{Odds}(\text{Kauf}) = \frac{P(\text{Kauf})}{1 - P(\text{Kauf})}$$

Exkurs: Vierfeldertafeln

Bei den Frauen beträgt die Chance auf Kauf also $0.401/0.599 = 0.67$, also in etwa 2:3.

Bei den Männern hingegen beträgt die Chance auf Kauf $0.257/0.743 = 0.346$ als in etwa 1:3.

Wenn man auch noch diese beiden Zahlen ins Verhältnis setzt, so nennt man das Ergebnis Chancenverhältnis oder Odds Ratio.

Hier besagt das Chancenverhältnis von $0.346/0.67 = 0.516$, daß die Chancen auf Kauf des Produkts bei Männern nur rund halb so groß sind wie bei Frauen.

Exkurs: Vierfeldertafeln

Das Chancenverhältnis kann auch leicht aus der Originaltabelle berechnet werden.

Geschlecht vs. Kauf	nein	ja
Frauen	273	183
Männer	627	217

$$\text{OddsRatio} \quad = \quad \frac{273 \cdot 217}{627 \cdot 183} \; = \; 0.516$$

Logistische Regression

Als Link-Funktion in der logistischen Regression wird die sogenannte Logit-Funktion

$$g(\mu) \quad = \quad \log\left(\frac{\mu}{1-\mu}\right)$$

verwendet. Diese berechnet die logarithmierten Chancen.

Damit ist also unser verallgemeinertes lineares Modell komplett:

$$\begin{aligned} E(y_i) &= \mu_i \\ g(\mu_i) &= x_i^\top \beta \end{aligned}$$

Logistische Regression

In der logistischen Regression ist dabei $g(\cdot)$ die Logit-Funktion und die Daten y_i sind binomialverteilt. Damit ist wieder die gesamte Likelihood spezifiziert und wieder kann der Parameter β über Maximum Likelihood geschätzt werden.

Bemerkung: Dasselbe Modell kann man im Allgemeinern *nicht* mit der Methode der kleinsten Quadrate schätzen.

Logistische Regression

Interpretation: Die logistische Regression beschreibt die Log-Odds

$$\log\left(\frac{P(y_i = 1 \,|\, x_i)}{P(y_i = 0 \,|\, x_i)}\right) \quad = \quad x_i^\top \beta$$

Beispiel: Betrachte den Zusammenhang von `choice` und `gender`. Die Odds der Männer sind dann

$$\frac{P(y = 1 \,|\, x = 1)}{P(y = 0 \,|\, x = 1)} \quad = \quad \exp(\beta_1 + \beta_2 \cdot 1)$$

Logistische Regression

Und die Odds der Frauen:

$$\frac{P(y=1\,|\,x=0)}{P(y=0\,|\,x=0)} \quad = \quad \exp(\beta_1 + \beta_2 \cdot 0)$$

und der Odds Ratio ist damit

$$\begin{aligned}\frac{\exp(\beta_1 + \beta_2 \cdot 1)}{\exp(\beta_1 + \beta_2 \cdot 0)} &= \frac{\exp(\beta_1 + \beta_2)}{\exp(\beta_1)} \\ &= \exp(\beta_1 + \beta_2 - \beta_1) \\ &= \exp(\beta_2)\end{aligned}$$

Logistische Regression

Somit gibt es also einen einfachen Zusammenhang zwischen den Koeffizienten der logistischen Regression und dem Odds Ratio.

Im Beispiel sind die Koeffizientenschätzer

$$\beta \quad = \quad (-0.400, -0.661)^\top$$

Die Chancen auf Kauf sind also bei Männern um rund 50% geringer: $\exp(-0.661) = 0.516$.

Tutorium

Verallgemeinerte Lineare Modelle in R: Logistische Regression (*GLM1.pdf*)

Übung

Aufgabe 8:

102 Psychologiestudenten der Universität Regensburg wurden gefragt, wie gut sie die Aussichten einschätzen nach ihrem Universitätsabschluß eine angemessene Anstellung bekommen: "erwarte keine adäquate Anstellung" (`"schlecht"`), "unsicher" (`"mittel"`) und "sofort nach dem Abschluß" (`"gut"`). Diese Daten sind im Datensatz *Job.rda* verfügbar und enthalten die Antworten in der Variablen `aussicht3`. Zur Analyse mit einem Modell für binäre Daten wurden die Kategorien `"schlecht"` und `"mittel"` in der Variablen `aussicht` zusammengelegt. Als erklärende Variable steht das Alter in Jahren (`alter`) zur Verfügung.

Übung

- Visualisieren Sie die Daten geeignet.
- Passen Sie ein logistisches Regressionsmodell an.
- Gibt es einen signifikanten Zusammenhang zwischen dem Alter und der Einschätzung der Jobaussichten?
- Um wieviel Prozent steigen/fallen die Chancen einer positiven Einschätzung durchschnittlich mit jedem Jahr?

Das verallgemeinerte lineare Modell

Das verallgemeinerte lineare Modell (GLM) ist gegeben durch:

$$\begin{aligned} E(y_i) &= \mu_i \\ g(\mu_i) &= x_i^\top \beta \end{aligned}$$

- y_i — Zufallsvariable aus einer Exponentialfamilie (bspw. normal-, binomial-, Poisson-verteilt) mit Erwartungswert μ_i.
- g — Link-Funktion, die von der Skala des Erwartungswertes μ_i auf die Skala des linearen Prädiktors (reelle Zahlen) abbildet.
- x_i — Vektor von Erklärungsvariablen.
- β — Vektor der zugehörigen Regressionskoeffizienten.

Das verallgemeinerte lineare Modell

Eine Variable y stammt aus einer Exponentialfamilie, wenn ihre Dichtefunktion ein Spezialfall folgender Dichte ist:

$$f(y \,|\, \theta, \phi) \quad = \quad \exp\left(\frac{y \cdot \theta - a(\theta)}{\phi} + b(y, \phi)\right)$$

wobei $a(\cdot)$ und $b(\cdot)$ bekannte Funktion sind, ϕ ein Skalenparameter ist (evtl. bekannt) und θ die Form der Verteilung kontrolliert.

Spezialfälle dieser Familie beinhalten:

- Normalverteilung
- Binomialverteilung
- Poissonverteilung

Das verallgemeinerte lineare Modell

Ist die konkrete Verteilung der y_i festgelegt, so ist wieder die komplette Likelihood der Beobachtungen spezifiziert und die Parameter β werden durch Maximum Likelihood (ML) geschätzt.

Da der ML-Schätzer (abgesehen von Spezialfällen) nicht geschlossen berechnet werden kann, wird er durch einen iterativen Algorithmus bestimmt. Dieser heißt IWLS (iteratively weighted least squares).

Das verallgemeinerte lineare Modell

Modellwahl:

Da die angepaßten GLMs mit ML geschätzt wurden, können zur Modellwahl sofort wieder t- bzw. F-Tests oder AIC bzw. BIC verwendet werden.

Für jedes GLM läßt sich leicht die maximierte log-Likelihood berechnen und damit auch das AIC bzw. BIC.

Die Formulierung der t- und F-Tests im linearen Modell hatten wir auf Basis der Fehlerquadratsummen vorgenommen, es gibt aber (fast) äquivalente Formulierungen auf Basis von maximierten log-Likelihoods, die im Falle von GLMs verwendet werden.

Das verallgemeinerte lineare Modell

Prognose:

Als Prognose für die abhängige Variable y_i wird, wie schon im linearen Modell, der geschätzte Erwartungswert $\hat{\mu}_i$ verwendet:

$$\hat{\mu}_i \quad = \quad g^{-1}\left(x_i^\top \hat{\beta}\right)$$

Dieser ist im GLM nicht direkt gleich dem linearen Prädiktor $x_i^\top \hat{\beta}$, sondern es wird zusätzlich noch die inverse Link-Funktion $g^{-1}(\cdot)$ angewendet.

Das verallgemeinerte lineare Modell

Residuen:

Als Residuen (oder Prognosefehler) im linearen Modell hatten wir einfach die Abweichung

$$y_i - \hat{\mu}_i$$

verwendet. Dies war möglich, da angenommen wurde, daß die Varianz aller Beobachtungen gleich ist.

Im GLM ist das im Allgemeinen nicht der Fall und deshalb gewichtet man die Residuen noch mit Hilfe ihrer Varianzen. Die Varianzfunktion $V(\mu)$ ist dabei immer durch die Wahl der Verteilungsfamilie festgelegt.

Das verallgemeinerte lineare Modell

Die Residuen

$$\frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

nennt man auch Pearson-Residuen. Die Varianzfunktion ist bspw.:

- normal: $V(\mu) = 1$,
- binomial: $V(\mu) = \mu \cdot (1 - \mu)$,
- Poisson: $V(\mu) = \mu$.

Es gibt aber auch andere sinnvolle Definitionen von Residuen in GLMs, bspw. Deviance-Residuen, die per Voreinstellung von R berechnet werden.

Das verallgemeinerte lineare Modell

Lineare Regression

Die lineare Regression ist ein Spezialfall des GLMs. Dabei wird angenommen, daß die y_i unabhängig normalverteilt sind mit derselben Varianz.

Als Link-Funktion verwendet man die Identität $g(\mu) = \mu$.

Dies entspricht dann genau den Annahmen (A1)–(A6).

Das verallgemeinerte lineare Modell

Binomialregression

Für binäre abhängige Variablen verwenden wir als Modell die Binomialverteilung. Wir haben dafür bereits die Logit Link-Funktion kennengelernt

$$g(\mu) \quad = \quad \log\left(\frac{\mu}{1-\mu}\right)$$

die gleichsam die logarithmierten Chancen modelliert.

Eine andere gängige Link-Funktion für binäre Daten ist die Probit-Funktion

$$g(\mu) \quad = \quad \Phi^{-1}(\mu)$$

wobei $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalvtlg. ist.

Das verallgemeinerte lineare Modell

Die Motivation dafür ist:

Nehmen wir an, wir möchten modellieren, ob ein Kunde ein bestimmtes Produkt testen möchte. Und weiters glauben wir, daß er das nur tut, wenn die Werbung für dieses Produkt bei ihm eine gewisse Reizschwelle überschreitet.

Die Reizschwelle R ist dabei nicht für alle potentiellen Kunden gleich sondern normalverteilt mit Erwartungswert m und Varianz s^2.

Das verallgemeinerte lineare Modell

Dann ist die Wahrscheinlichkeit p, daß er das Produkt probiert gegeben als

$$p \quad = \quad P(R \leq \text{Werbung}) \quad = \quad \Phi\left(\frac{\text{Werbung} - m}{s}\right)$$

und damit gilt

$$\Phi^{-1}(p) \quad = \quad -\frac{m}{s} + \frac{1}{s} \cdot \text{Werbung} \quad = \quad \beta_1 + \beta_2 \cdot \text{Werbung}$$

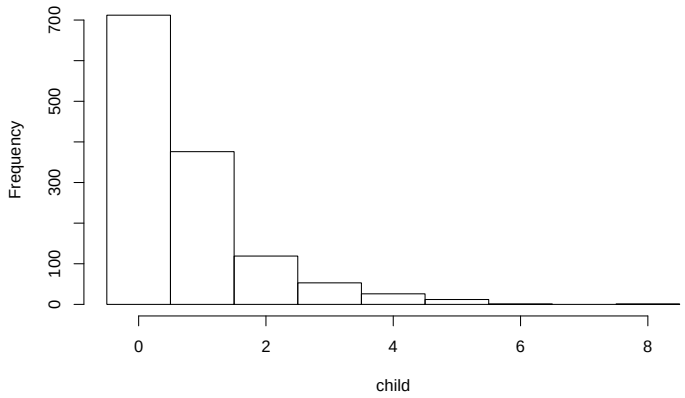
Dies entspricht genau einem GLM mit Probit-Link.

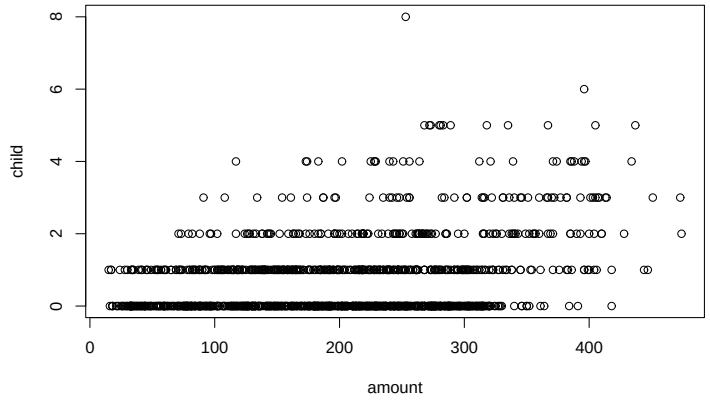
Poisson-Regression

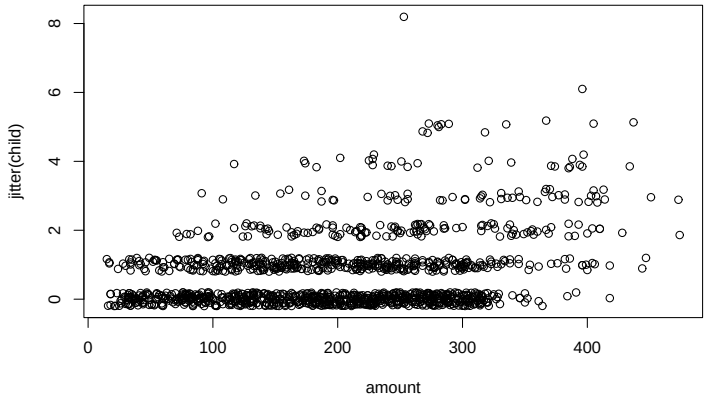
Um Zähldaten zu modellieren wird häufig die Poisson-Verteilung verwendet.

Als Link-Funktion wird dann in aller Regel der Logarithmus eingesetzt: $g(\mu) = \log(\mu)$

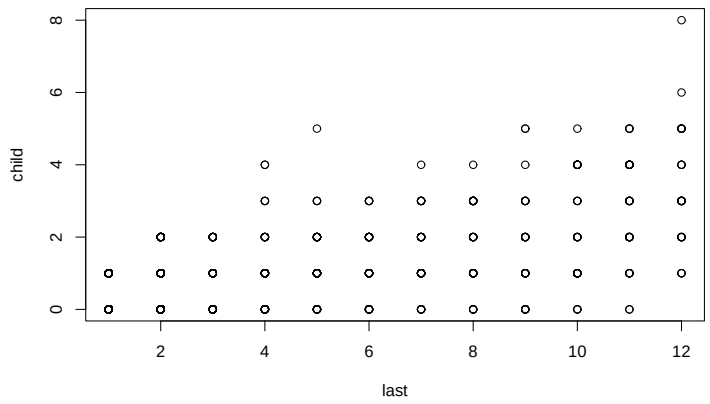
Beispiel: Wie hängt die Anzahl der gekauften Kinderbücher `child` von den verfügbaren Erklärungsvariablen im `BBBClub` Datensatz ab?



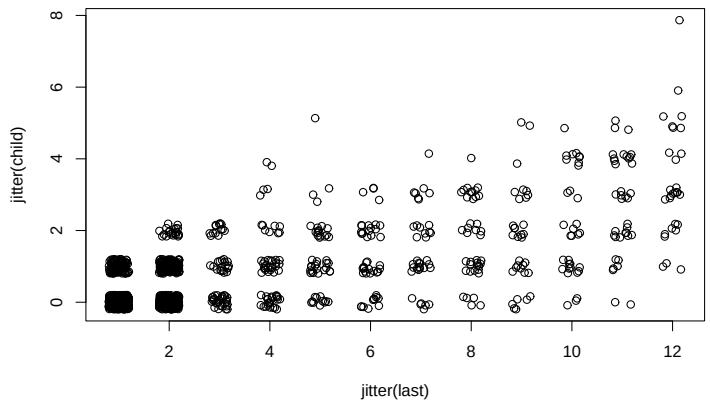




Das verallgemeinerte lineare Modell



Das verallgemeinerte lineare Modell



Tutorium

Verallgemeinerte Lineare Modelle in R (*GLM2.pdf*)

Übung

Aufgabe 9:

Bei Kreditvergaben sind Banken daran interessiert, ob neue Kunden den Kredit, den sie aufnehmen, auch wirklich zurückzahlen. Das Ziel von Kredit-Scoring ist es, die Wahrscheinlichkeit für die Rückzahlung eines Kredits auf Basis von bestimmten Risikofaktoren des Kunden zu modellieren bzw. vorherzusagen. Der Datensatz *Kredit.rda* enthält die Daten über 1000 Konsumkredite, die von einer deutschen Bank vergeben wurden.

Wie hängt die Rückzahlungswahrscheinlichkeit von den verfügbaren Erklärungsvariablen ab?

Übung

- `kredit` Wurde der Kredit zurückgezahlt?
- `konto` Hat der Kunde auch ein Sparkonto? Die Variable hat drei Kategorien "kein Konto" (`"kein"`), "mittelmäßiges Konto" (weniger als DEM 200, `"mittel"`, Referenzkategorie), "gutes Konto" (`"gut"`).
- `laufzeit` Laufzeit des Kredits in Monaten.
- `hoehe` Höhe des Kredits in DEM.
- `moral` Zahlungsmoral bei bisherigen Krediten.
- `verwend` Verwendungszweck des Kredits: professionell oder privat.
- `geschlecht` Geschlecht des Kreditnehmers.

Übung

Aufgabe 10:

Um die Effektivität ihrer TV-Spots zu untersuchen, hat eine Firma erhoben, von wie vielen Personen sie Reaktionen auf den Spot bekommen haben. Der Datensatz $TVSpot.rda$ enthält wöchentliche Informationen über etwas mehr als drei Jahre:

- `pers` Anzahl von Reaktionen.
- `woche` Nummer der Woche.
- `aufwand` Finanzieller Aufwand für die Werbung.

Wie verändert sich die Zahl der Reaktionen mit der Zeit und dem Werbeaufwand?

Andere Regressionsmodelle

Choice-Modelle

- **Multinomialmodell**: Ähnlich dem Binomialmodell, aber die abhängige Variable kann $c > 2$ Ausprägungen haben.
- **Ordinale logistische Regression**: oder proportional odds model. Die abhängige Variable hat $c > 2$ geordnete Ausprägungen (bspw. Rating).
- **Conjoint Analyse**: Ähnlich dem Multinomialmodell, aber nicht jedes Individuum hat alle c Möglichkeiten zur Auswahl.

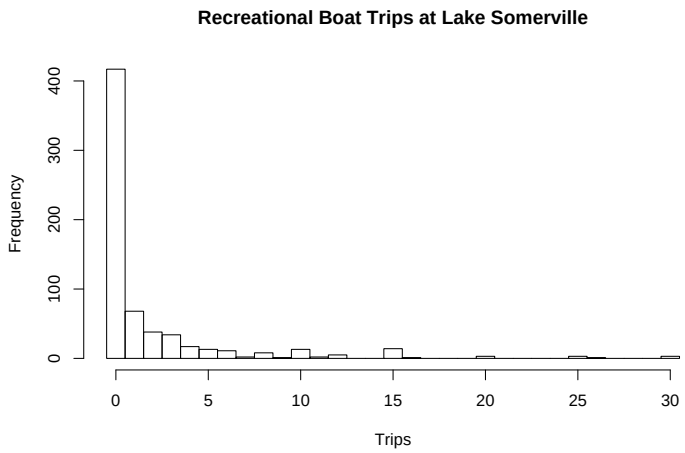
Andere Regressionsmodelle

Zähldatenmodelle

In vielen ökonomischen Datensätzen von Zähldaten kommen mehr 0-Beobachtungen vor als durch viele Verteilungen modelliert werden. Dies wird dann oft speziell behandelt, bspw.

- Zero-inflated Poisson (ZIP) Modell,
- Hürdenmodell.

Andere Regressionsmodelle



Andere Regressionsmodelle

In vielen Anwendungen ist man nicht an einer probabilistischen Modellierung der Daten interessiert, sondern lediglich an einem Algorithmus, der als "Prognosemaschine" benutzt werden kann. Hier gibt es zahlreiche Algorithmen aus dem **maschinellen Lernen** bzw. **statistischen Lernen**. Bspw.:

- Neurale Netze,
- Support Vector Machine,
- Baumbasierte Verfahren: Klassifikationsbäume, Bagging, Random Forest,
- Boosting

Andere Regressionsmodelle