Lineare und logistische Regression

LE1

Dr. Lucia Kleint

Fachhochschule
Nordwestschweiz

Herbstsemester 2021

Inhaltsverzeichnis

Ziel der Kompetenz

Die Statistik ist die Grundlage zur zuverlässigen Auswertung von Daten. Allerdings sind Daten nie perfekt und somit muss ein geeignetes Modell gefunden werden. Das Hauptziel dieser Kompetenz ist zu verstehen, wie Daten in der Data Science durch lineare und logistische Regression praktisch ausgewertet werden können.

Jedes Teil-Skript enthält ein Lernergebnis, inklusive Verweisen auf detailliertere Quellen wie Bücher. Für die Prüfung werden die Themen dieses Skripts, aber so detailliert wie in den Verweisen vorausgesetzt.

last update: 30. September 2021

Verweise

Folgende Bücher/Unterlagen werden für das genauere Verständnis empfohlen:

- L. Fahrmeir et al., Statistik, Der Weg zur Datenanalyse, 8. Auflage, Springer (pdf via FHNW Netzwerk). Kapitel 3.6
- wst Skript, Prof. Dr. M. Steiner-Curtis, 2015, Kap. 10 und 11.
- Skript Applied Statistics and Data Analysis, Prof. Dr. M. Steiner-Curtis, 2020, Kap. 7 und 8
- Video D. Jung https://www.youtube.com/watch?v=WZKMzjfJ-x4
- Video D. Jung https://www.youtube.com/watch?v=zsEn-oqSVhY
- Video M. Kunter https://www.youtube.com/watch?v=osPwkFkYvNc

Section 1

Repetition

Repetition

Satz (Arithmetisches Mittel)

Das arithmetische Mittel $\bar{x}$ von n Datenpunkten $x_1, ..., x_n$ ist

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Satz (Median)

Der Median ist für ungerades n der mittlere Datenpunkt der geordneten Liste und für gerades n das arithmetische Mittel der beiden in der Mitte liegenden Datenpunkte.

$$x_{med} = \begin{cases} x_{(\frac{n+1}{2})} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) \end{cases}$$

Repetition

Satz (Varianz)

Die Varianz der Werte $x_1, ..., x_n$ ist

$$\sigma^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Satz (Standardabweichung)

Die Standardabweichung ist die Wurzel aus der Varianz.

$$\sigma = +\sqrt{\sigma^2}$$

Satz (Standardfehler des Mittelwerts)

Der Standardfehler des Mittelwerts ist

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Section 2

Lineare Methoden für Regression und Klassifikation

Beispiel zur Einleitung

Die folgende Grafik zeigt das Gehalt bei der Firma ABC abhängig von den Jahren an Berufserfahrung.

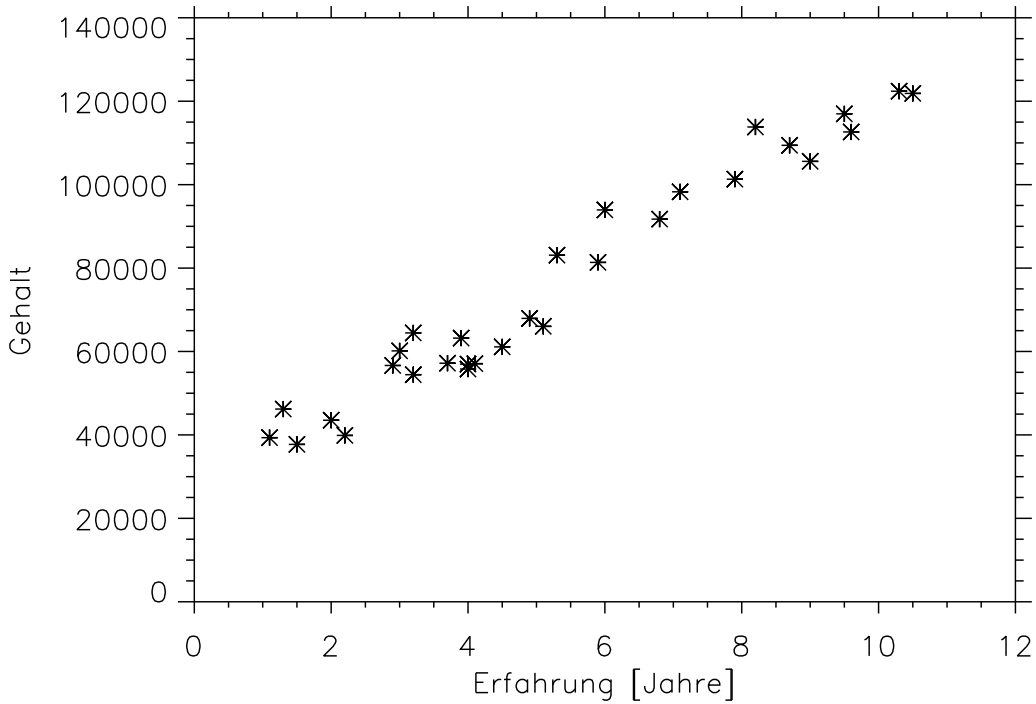


Abbildung: Daten von: https://s3.us-west-2.amazonaws.com/public.gamelab.fun/dataset/salary_data.csv

Als Data Scientist sollen Sie ein Computerprogramm schreiben, welches der HR Abteilung hilft, das Gehalt bei zukünftigen Einstellungen schnell festzulegen. Wie gehen Sie vor?

Beispiel zur Einleitung

Man sieht sofort, dass die Punkte mehr oder weniger auf einer geraden Linie liegen. Man wird also ein Modell versuchen, welches so aussieht:

$$\text{Gehalt} = \alpha \cdot \text{Erfahrung} + \beta$$

Wie man das Modell bestimmt, welche Genauigkeit und Fehler es hat, ist das Ziel dieses Lernergebnisses. Sie werden lernen, das folgende Modell zu bestimmen.

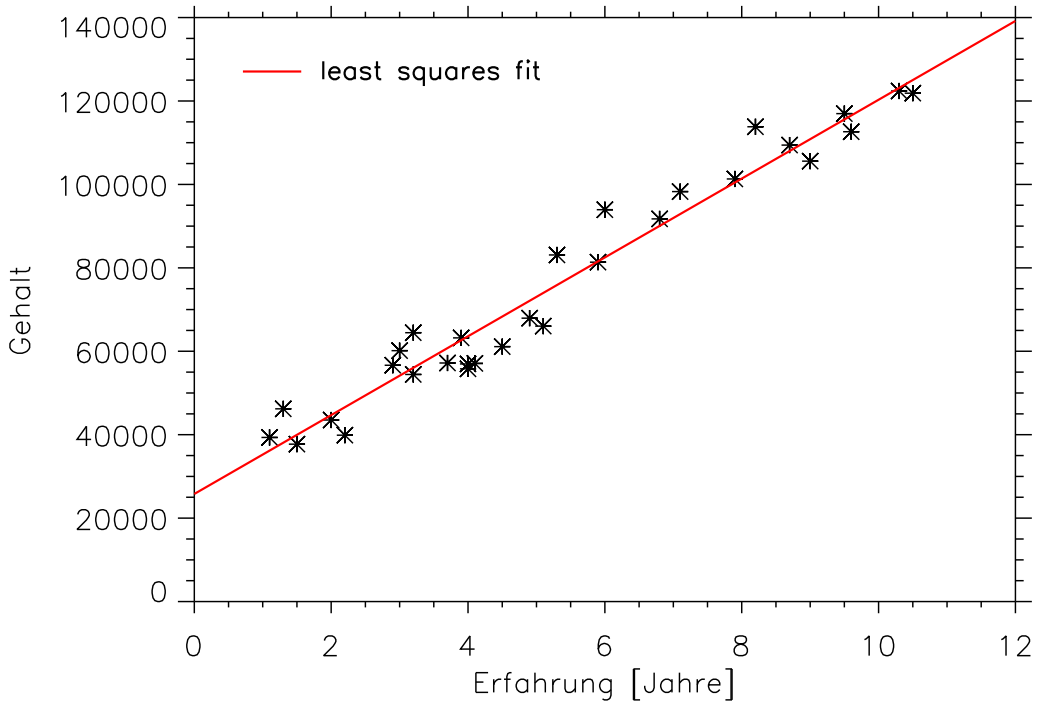


Abbildung: Lineare Regression

Lineare Zusammenhänge

Ein Zusammenhang zwischen zwei Messgrössen kann als $Y = f(X)$ geschrieben werden und das Ziel ist, die Funktion f möglichst gut zu bestimmen. Allerdings sind die Datenpunkte oft noch durch andere Grössen beeinflusst, was man als Fehlerterm schreiben kann: $Y = f(X) + \epsilon$. Diese Art von Problemen wird Regression genannt.

Eine erste Näherung für korrelierte Datenpunkte (z.B. Gehalt vs. Erfahrung oder Mietpreis vs. Wohnungsgrösse) wird oft durch eine lineare Funktion versucht, also eine Funktion der Form $f(X) = \alpha + \beta X$. Für die Datenpaare (x_i, y_i) gilt dann die lineare Beziehung $y_i = \alpha + \beta x_i + \epsilon_i$, wobei ϵ_i den Fehler bezeichnet. Im Folgenden werden wir annehmen, dass die Fehler ϵ_i normalverteilt sind. Dies nennt man lineare Einfachregression, das einfachste und bekannteste Regressionsmodell.

kategoriale und kontinuierliche Variablen

Wir werden mit zwei Arten von Variablen arbeiten: kategoriale und kontinuierliche.

- Kategoriale Variablen sind Merkmale, die in Kategorien unterteilt werden und eine begrenzte Anzahl von Kategorien haben. Beispiele: Familienstand, Einkommensklasse, Kinderzahl.
- Kontinuierliche Variablen haben sehr viele Ausprägungen und kontinuierliche Eigenschaften. Beispiele: Einkommen, Ausbildungsdauer

kategoriale und kontinuierliche Variablen

Bei kategoriale Variablen modelliert man das Auftreten der einzelnen Kategorien (z.B.: Wie gross ist die Wahrscheinlichkeit, verheiratet zu sein). Bei kontinuierlichen Variablen ist dies nicht praktikabel. Stattdessen analysiert man bestimmte Eigenschaften (Zentrum, Streuung) der Verteilung aller Ausprägungen (z.B.: Wie hoch ist das Durchschnittseinkommen?).

Ab welcher Anzahl von Ausprägungen man eine Variable nicht mehr als kategoriale Variable betrachtet, ist dabei nicht festgelegt und hängt sowohl von inhaltlichen als auch von praktischen Gesichtspunkten ab.

Je nach Variablentyp benützt man andere Modelle.

Abhängige Variable	**Unabhängige Variablen**	
	kontinuierlich	**kategorial**
kontinuierlich	Lineare Regression	Lineare Regression mit Dummies
kategorial (dichotom)	Logistische Regression	Logistische Regression mit Dummies
kategorial (polytom)	Multinomiale Regression	Multinomiale Regression mit Dummies

Abbildung: Quelle: http://eswf.uni-koeln.de/lehre/06/03/ss0603_11.pdf

Wenn man Datenpunkte hat, die nicht perfekt linear sind, dann ist die Frage, welcher Fit am besten geeignet ist. Im folgenden Beispiel ist dies veranschaulicht: Welche der drei farbigen Geraden ist der beste Fit?

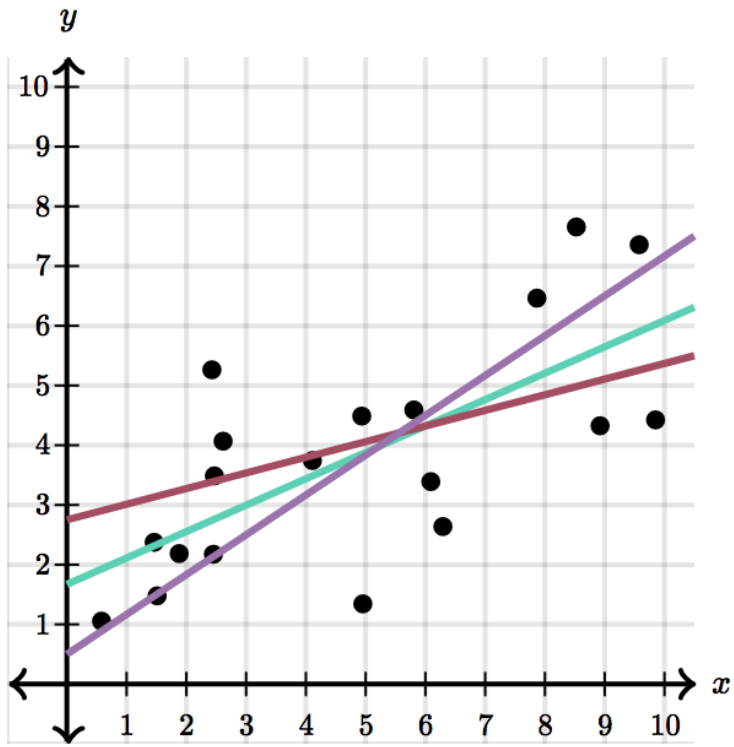


Abbildung: Quelle: https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/regression-library/a/introduction-to-residuals

Die Lösung ist, dass die Abweichung der Geraden zu den Datenpunkten minimal sein soll, was man durch eine mathematische Prozedur bestimmen kann. Diese nennt man die Methode der kleinsten Quadrate.

Lineare Regression - Kleinste Quadrate

Das Ziel der linearen Regression ist die Koeffizienten α und β, also den Achsenabschnitt und die Steigung zu bestimmen, sodass die Datenpunkte möglichst wenig von der Gerade entfernt liegen. Diese Abweichung wird global minimiert. Üblicherweise wird die Summe der quadrierten Differenzen zwischen den Messwerten und den Funktionswerten in Abhängigkeit von α und β minimiert:

$$Q(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i')^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2,$$

wobei y_i' die Funktionswerte und y_i die Messwerte bezeichnen. Die Minimierung von $Q(\alpha, \beta)$ bezeichnet man als Methode der kleinsten Quadrate.

Lineare Regression - Kleinste Quadrate

Grafisch kann man es sich folgendermassen vorstellen. Wenn man die grünen Datenpunkte hat, dann ist die blaue Linie der beste Fit, der die Quadrate der Fehler (rote Pfeile) minimiert.

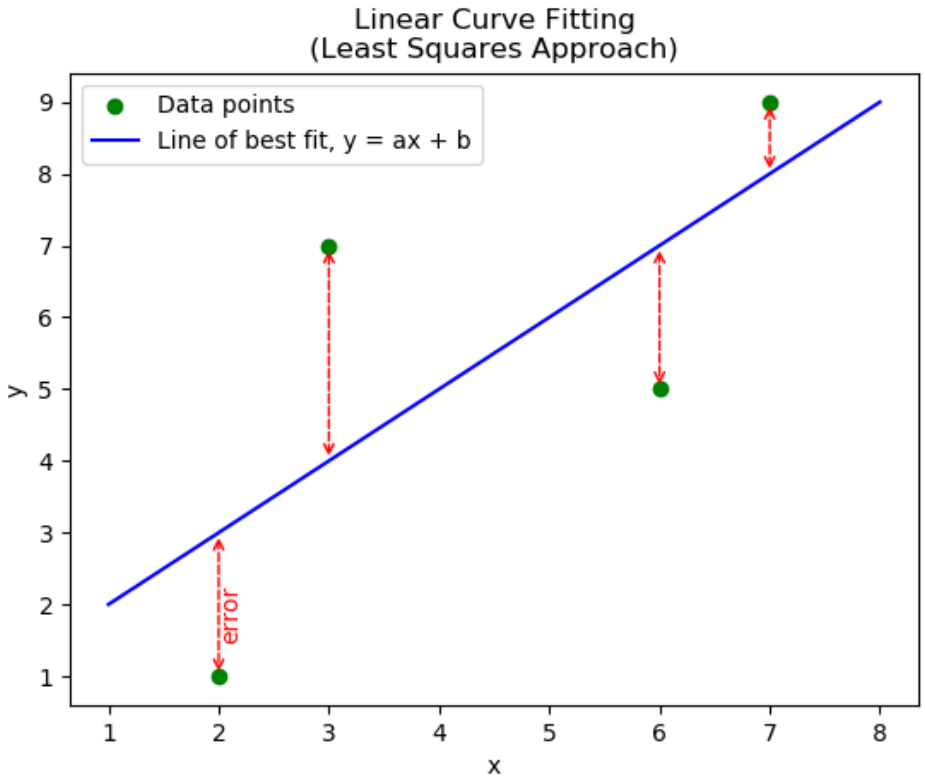


Abbildung: Quelle: https://blog.mbedded.ninja/mathematics/curve-fitting/linear-curve-fitting/

Kleinste Quadrate

Die Minimierung von $Q(\alpha, \beta)$ erhält man, indem man die Funktion partiell nach α und β differenziert und gleich null setzt. Die Schritte können in Fahrmeir's Buch auf Seite 146 nachgelesen werden oder z.B. hier: https://de.wikipedia.org/wiki/Lineare_Einfachregression.

Man findet die Ausdrücke

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

und

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

wobei $\bar{x}$ und $\bar{y}$ die Mittelwerte der gemessenen Daten sind. Die Gerade $\alpha + \beta$x wird Regressionsgerade genannt.

Kleinste Quadrate

Bei dieser Methode muss man allerdings aufpassen, dass man keine Ausreisser im Datensatz hat, was man am besten grafisch verifiziert. Im folgenden Beispiel ist dargestellt, dass in einem solchen Fall (der Ausreisser ist der weisse Punkt) die Gerade falsch herauskommen kann, oder man fälschlicherweise eine Abhängigkeit findet, obwohl es keine gibt.

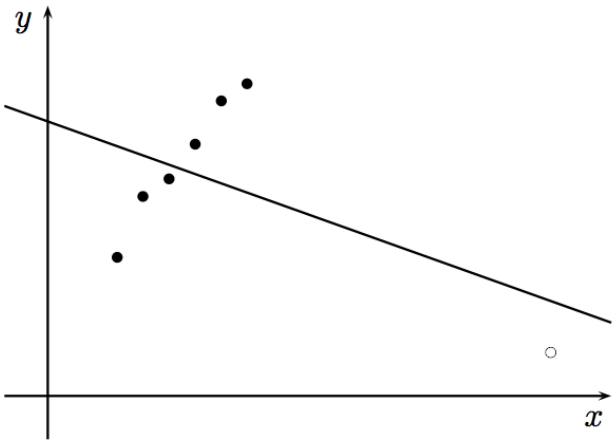


Abbildung 10.1.ii: Eine falsche Regressionsgerade wegen einem Ausreisser

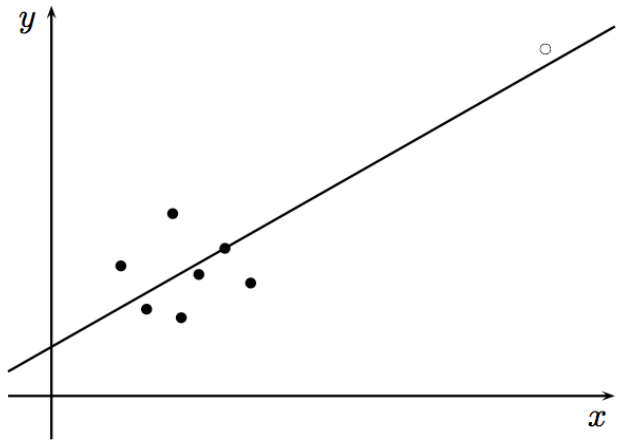


Abbildung 10.1.iii: Eine vorgetäuschte Abhängigkeit wegen einem Ausreisser

Abbildung: Quelle: wst Skript, Prof. Steiner-Curtis, 2015

Kleinste Quadrate

Oft ist die Matrixschreibweise praktisch, vor allem wenn man dieses Problem im Computer effizient lösen will: Wir lösen die lineare Gleichung $Ax = b$, wobei $x = [\beta, \alpha]^T$, $b = [y_1, y_2, ..., y_n]^T$, und $A = [x_1, 1; x_2, 1; ...; x_n, 1]$.
x ist dann $x = (A^T A)^{-1} A^T b$.

Die Abweichung zwischen den Messwerten und den gefitteten Werten nennt man Residuen und bezeichnet sie mit $\epsilon_i' = y_i - y_i'$. Es ist oft hilfreich, die Residuen zu plotten. Wenn eine systematische Abweichung zu erkennen ist, war das lineare Modell wahrscheinlich keine gute Annahme.

Wir nehmen hier implizit an, dass $(A^T A)$ invertierbar oder nicht-singulär ist. Falls dies nicht der Fall ist, dann ist die Lösung des Problems der kleinsten Quadrate nicht eindeutig (und man müsste kompliziertere Methoden verwenden).

Residuen

Die Residuen sind die Differenz zwischen dem Messwert und dem Fit (Beispiele in rot). Es ist sinnvoll, die Residuen jeweils zusammen mit dem Fit zu plotten und zu schauen, wie ihre Streuung aussieht. Im Fall eines guten Fits streuen die Residuen um 0 (linker Plot).

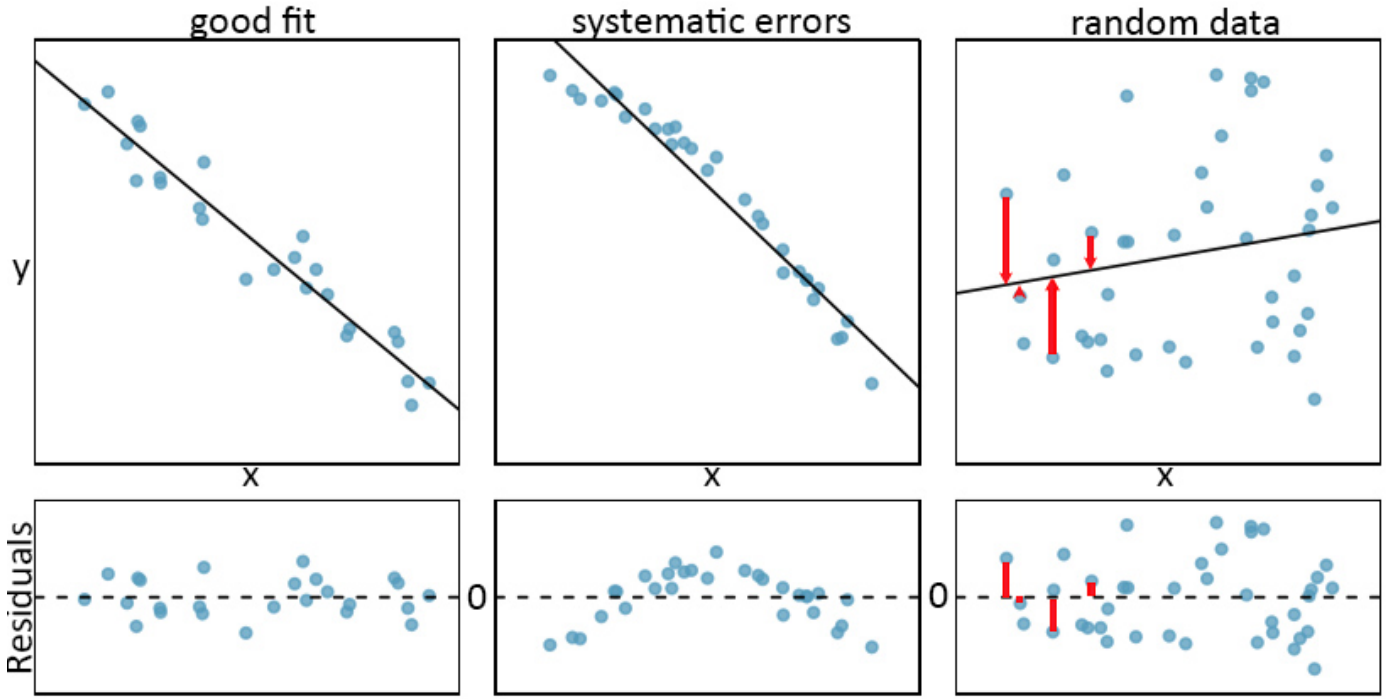


Abbildung: Im mittleren Plot sieht man eine systematische Abweichung (eher quadratische Funktion) und deswegen ist ein linearer Fit nicht ideal. Im rechten Plot sind die Datenpunkte mehr oder weniger zufällig verteilt und deswegen streuen auch die Residuen sehr. In diesem Fall macht ein Geradenfit keinen Sinn, da keine lineare Beziehung zwischen x und y besteht. Angepasst von https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_OpenIntro_Statistics_(Diez_et_al)./07%3A_Introduction_to_Linear_Regression/7.02%3A_Line_Fitting%2C_Residuals%2C_and_Correlation

Residuenanalyse

Der erste Test, den man bei linearen Modellen durchführen sollte ist die Residuenanalyse. Mit dieser Analyse interpretiert man die Grafiken der Residuen des Modells und kann oft bereits einfach feststellen, wenn ein Modell nicht passt.

Die verwendeten Grafiken heissen Tukey-Anscombe-Diagramm. Statt der Originaldaten (x_i, y_i) plottet man einfach (x_i, r_i), nämlich die Werte der Residuen r_i auf der y-Achse. In diesem Diagramm sollten die Residuen gleichmässig um r=0 streuen.

Wir prüfen dabei die folgenden Annahmen: 1) Die Residuen haben den Erwartungswert 0, d.h. es sollte keine systematischen Fehler geben im Modell. 2) Die Fehler sollten unabhängig voneinander sein. 3) Die Fehler sollten normalverteilt sein, was man nicht im Tukey-Anscombe-Diagramm prüft, sondern in einem Histogramm.

Residuenanalyse

Die folgenden Beispiele zeigen Tukey-Anscombe-Diagramme und testen die Annahmen 1) und 2).

Abbildung 5.5: Tukey-Anscombe Plots. (a) Der Plot zeigt einen Trichter, der nach rechts grösser wird. D.h., die Fehlervarianz nimmt zu, wenn die beobachteten Werte der Zielgrösse gross sind. Damit ist die Annahme der gleichen Fehlervarianzen verletzt und die lineare Regression liefert keine zuverlässigen Werte. Evtl. kann man das Problem lösen, indem man die Zielgrösse transformiert. (b) Der Plot zeigt einen Trichter, der nach rechts kleiner wird. Wie in Fall (a) sind die Annahmen verletzt und die Regression liefert keine zuverlässigen Schätzwerte. (c) Die Punkte folgen einer U-förmigen Kurve. Das wahre Modell scheint komplizierter zu sein als das von uns angepasste. Evtl. kann man das Problem lösen, indem man einen quadratischen Term als erklärende Variable hinzufügt. (d-i) In diesen Plots ist kein Muster zu erkennen. Sie stellen einwandfreie Tukey-Anscombe Plots dar.

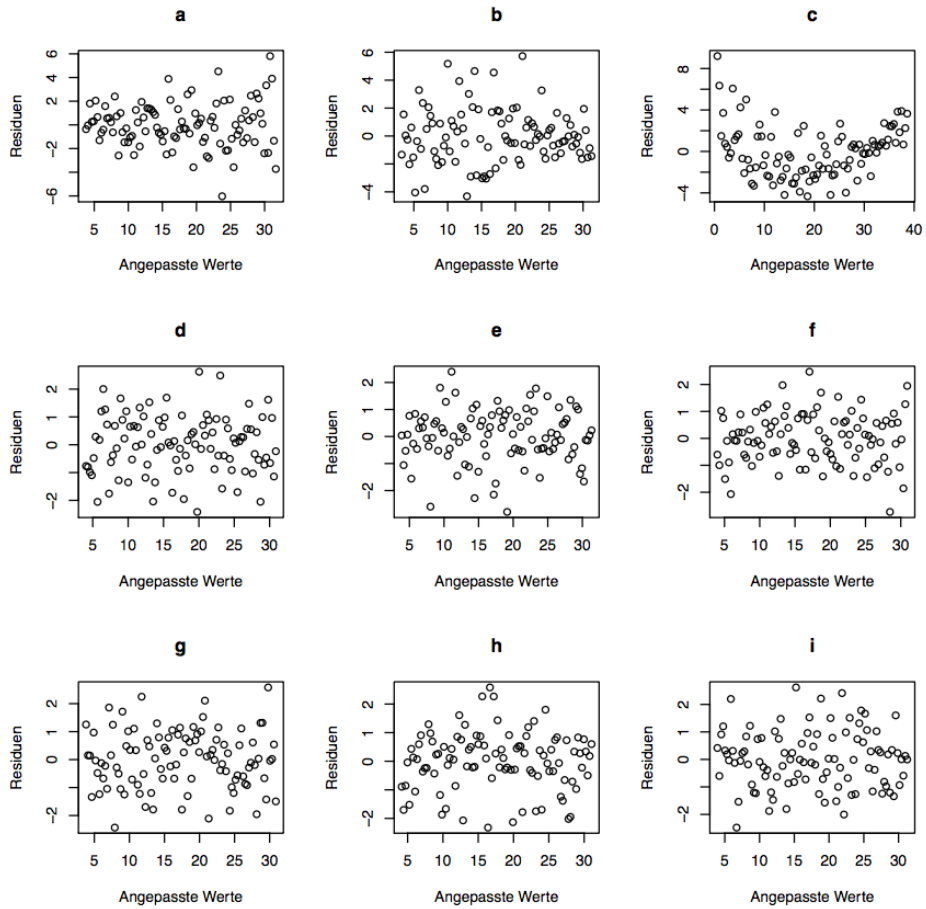


Abbildung: Quelle: Skript Einführung Statistik, ETH, Weiterbildungs-Lehrgang 2019-2021, M. Kalisch et al.

Residuenanalyse

Die Annahme 3) sagt, dass die Residuen normalverteilt sind, was man durch Fitten einer Normalverteilung prüfen kann. Im folgenden Beispiel wird ein Fall gezeigt, wo es relativ gut stimmt. Natürlich ist es wieder abhängig vom Problem, welche Abweichungen man toleriert, z.B. ob hier die Spitze bei -0.15 signifikant ist oder nicht.

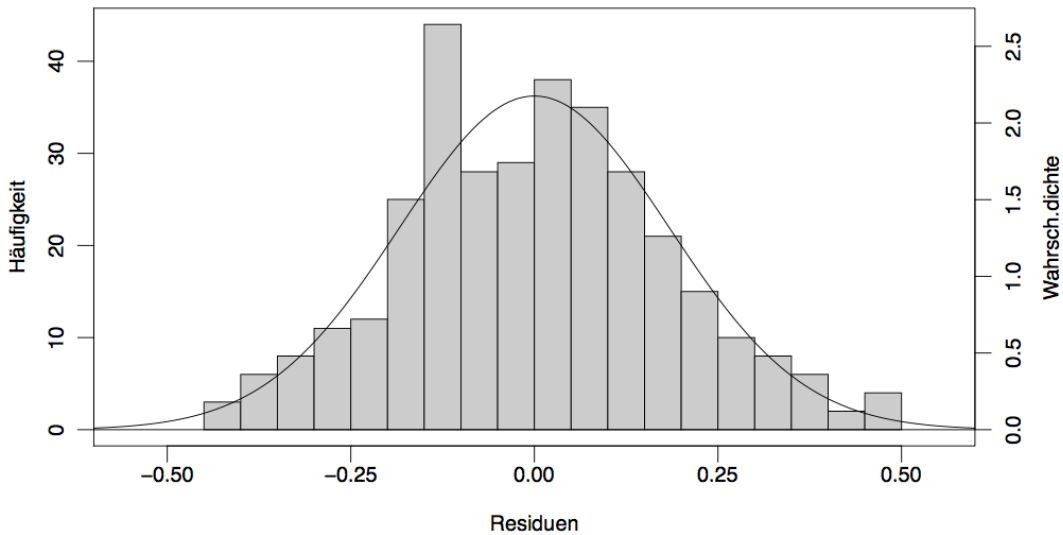


Abbildung: Quelle: Skript Statistische Regressionsmodelle, ETH, W. Stahel, 2017

Residuenanalyse

Wenn die Residuen nicht Annahmen 1)-3) erfüllen, muss man ein besseres Modell wählen. Man kann folgende Verbesserungen versuchen

- Variable transformieren (z.B. x_i^2 statt x_i verwenden)
- zusätzliche Terme ins Modell aufnehmen (wird in der Kompetenz mlr vorkommen)
- allgemeinere Modelle und statistische Methoden verwenden (die erst in späteren Kompetenzen vorkommen)

Beispiel für lineare Regression in Python

Lineare Regression kann gut mit scikit-learn in Python durchgeführt werden. Die Kurzfassung ist, dass man in Python ein Modell erstellt, via

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```

und dann den Fit durchführt mittels

```
model.fit(x, y)
```

Das Bestimmtheitsmass (siehe nächste Folie) erhält man mit
`r_sq = model.score(x, y)`

Die folgende Webseite hat ein gute Erklärung zum Vorgehen mit Python:
https://realpython.com/linear-regression-in-python/

Bestimmtheitsmass R^2

Bei einem guten Modell sind die Residuen klein. Allerdings ist dies noch kein quantitatives Kriterium zur Modellgüte. Das sogenannte Bestimmtheitsmass R^2 ist das Verhältnis der erklärten Quadratsumme (SQE) zur totalen Quadratsumme (SQT), mathematisch

$$R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n}(y_i' - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

wobei y_i die gemessenen Werte sind, y_i' die gefitteten, und $\bar{y}$ der Mittelwert der gefitteten Werte $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i'$.
Man kann die obige Formel auch umformen zu

$$R^2 = \frac{\sum_{i=1}^{n}(y_i' - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Bestimmtheitsmass R^2

Das Bestimmtheitsmass zeigt, wie gut das durch die Schätzung gefundene Modell zu den Daten passt. Wenn alle Residuen $= 0$ sind, d.h. die Datenpunkte alle perfekt auf einer Geraden liegen, dann ist $R^2 = 1$. Für den Fall von Datenpunkten auf einer Horizontalen (z.B. x=$\{1,2,3\}$ und y=$\{3,3,3\}$), hätte man in der Formel $0/0$, was undefiniert ist und somit wäre in so einem Spezialfall R^2 undefiniert. Im Normalfall ist aber $0 \leq R^2 \leq 1$, wobei je näher das Bestimmtheitsmass bei 1 ist, desto kleiner ist die Residuenquadratsumme.

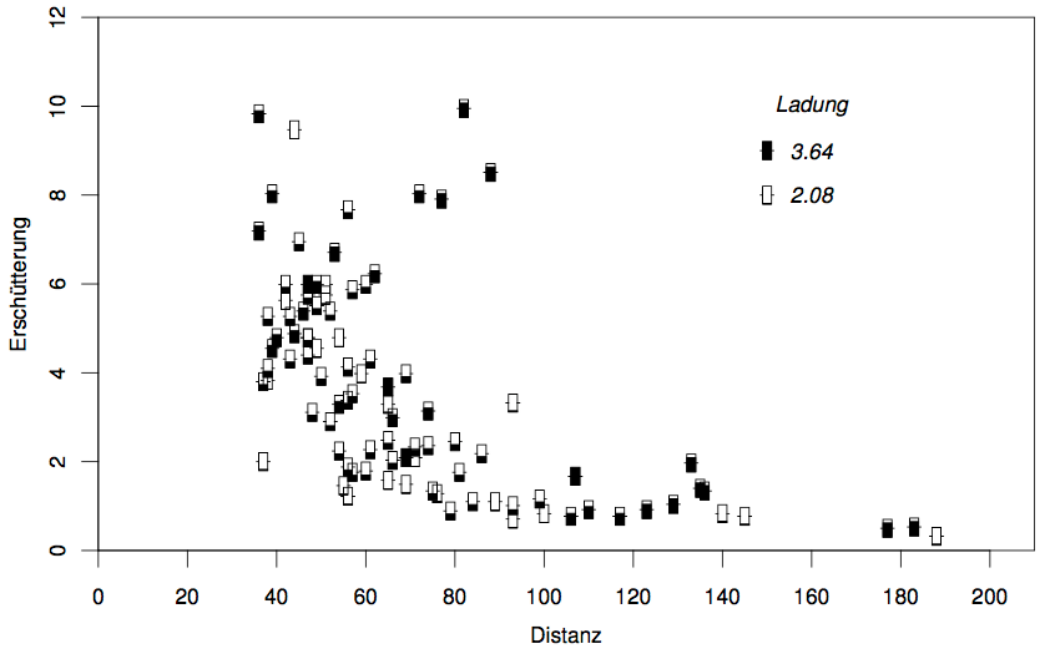
Der Grund für die Schreibweise als R^2 ist, dass in der linearen Regression das Bestimmtheitsmass das Quadrat des Bravais-Pearson-Korrelationskoeffizienten ist.

kleinste Quadrate: Beispiel

Beispiel

Beim Bau eines Tunnels muss gesprengt werden. In der Nähe von Häusern darf die Erschütterung dabei nicht zu hoch sein. Wir suchen nun das Gesetz, welche Sprengladung wo angewendet werden darf.Beispiel aus Skript von W. Stahel, ETH Skript, 2013

Die Erschütterung ist abhängig von der Ladung, der Distanz zwischen Spreng- und Messort, vom Untergrundmaterial, etc., wobei einige dieser Grössen unbekannt sind (z.B. der Untergrund).



kleinste Quadrate Beispiel

In der vorherigen Grafik sieht man, dass eine Gerade kein guter Fit wäre. Wenn man nun allerdings die Variablen transformiert, in diesem Fall mit einem Logarithmus, dann sieht es schon besser aus.

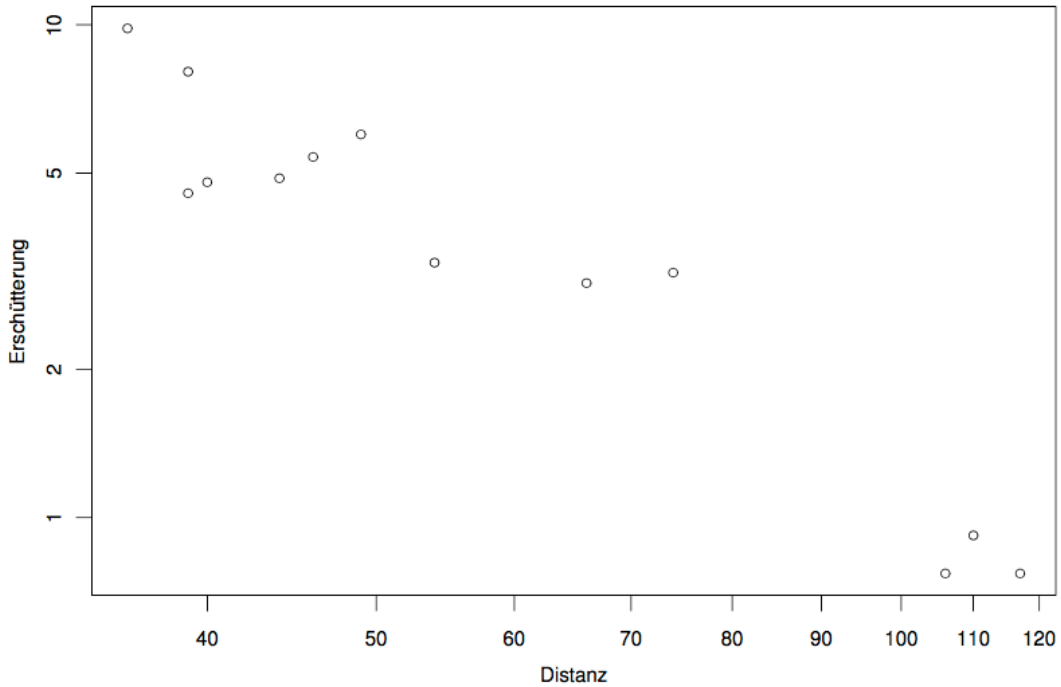
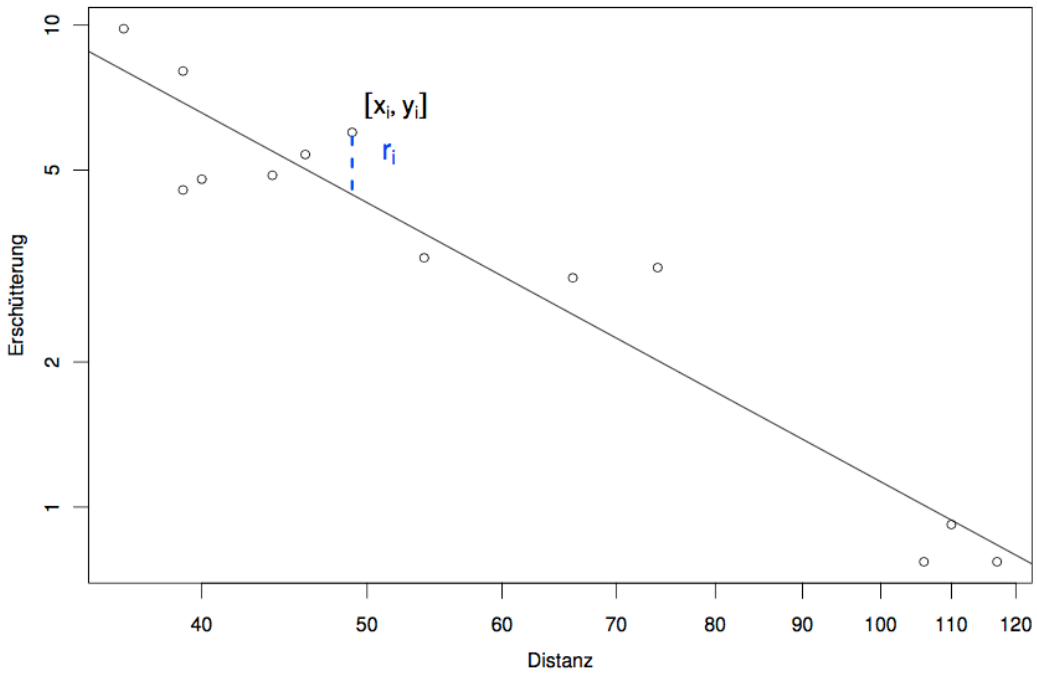


Abbildung: Die Achsen sind nun logarithmisch dargestellt und nur die Datenpunkte mit konstanter Sprengladung 3.12 sind eingezeichnet.

Das Transformieren der Daten ist immer erlaubt. Das "linear" in "lineare Modelle" bezieht sich auf die Koeffizienten, nicht auf das physikalische Gesetz. Man kann somit immer $y = a + bx$ umformen zu $\log(y) = c + d\log(x)$ und durch Ersetzen der Variablen (z.B. $\log(y) = q$) daraus wieder eine lineare Gleichung machen ($q = c + d \cdot r$).

kleinste Quadrate Beispiel

In diesem Fall wird aus der Physik erwartet, dass die Erschütterung proportional ist zu einer Konstanten · Ladung/(Distanz)^2. Durch Transformation erhält man somit log(Erschütterung) prop. log(const) + log(Ladung) - 2 · log(Distanz), was ein linearer Zusammenhang ist und deswegen erwartet man die Datenpunkte auf einer Geraden.



Man bestimmt die Parameter α und β mittels der Gleichung für die kleinsten Quadrate und findet die eingezeichnete Gerade. Dabei wurden die quadrierten Distanzen r_i^2 minimiert.

Logistische Regression

Bisher war die abhängige Variable (y) immer eine kontinuierliche Variable. Aber was wenn man nun ein Problem hat, wo y nur zwei Werte annimmt, z.B. der Einfluss des Einkommens darauf, ob der Haushalt ein Auto hat ($y = 1$) oder nicht ($y = 0$). Dies ist ein häufiger Fall, denn oft interessiert man sich in der Praxis nur für ja/nein Antworten (Kreditwürdigkeit?; Wird ein Produkt gekauft durch Marketing?; Wird eine Partei gewählt?; ...).

Solche Aufgaben löst man mit der logistischen Regression. Die lineare Regression eignet sich hier nicht, da ihre Voraussetzungen, z.B. die Normalverteilung der Residuen, nicht gegeben sind und sie auch generell nicht nur Werte zwischen 0 und 1 liefert. Die Methode nennt man dichotome logistische Regression, wobei sich dichotom auf die zwei Möglichkeiten (ja/nein) bezieht. Den Fall der multinomialen logistischen Regression (mit mehr als 2 Optionen für Antworten) werden wir hier nicht behandeln.

Logistische Regression

Die Zielvariable y kann als Wahrscheinlichkeit formuliert werden:

$$\pi_i = P(y_i = 1), \qquad 1 - \pi_i = P(y_i = 0).$$

mit der Auftretenswahrscheinlichkeit π_i (z.B. Wahrscheinlichkeit für Autobesitz), die natürlich von den Inputwerten (z.B. Einkommen, Bildung, Familienstand) abhängt. $\pi_i \in [0, 1]$, da es eine Wahrscheinlichkeit ist. Dies kann man mit linearer Regression nicht fordern, aber im logistischen Regressionsmodell transformiert man den Erwartungswert so, dass es zutreffen muss. Das logistische Regressionsmodell lautet

$$P(y_i = 1) = \pi_i = \frac{\exp\left(\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p\right)}{1 + \exp\left(\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p\right)} = \frac{\exp\left(x_i^T\beta\right)}{1 + \exp\left(x_i^T\beta\right)}$$

mit den Regressionskoeffizientenvektor β. Diese Gleichung ist der allgemeine Fall, in dieser Lerneinheit setzen wir alle Terme ab $\beta_2 = 0$, da wir nur die einfache Regression betrachten.

Logistische Regression

Beispiel

Ein Video, welches das Prinzip der logistischen Regression erklärt, kann hier angeschaut werden https://www.youtube.com/watch?v=yIYKR4sgzI8

Das wichtigste, was man sich merken sollte, ist dass die Wahrscheinlichkeit immer zwischen 0 und 1 liegt. Dabei steht 0 für die eine Option (im Video "normalgewichtig") und 1 für die andere Option (im Video "fettleibig").

Logistische Funktion

Satz

Die logistische Funktion ist eine stetige Wahrscheinlichkeitsverteilung. Sie modelliert die Grösse einer Population P zum Zeitpunkt t als

$$P(t) = \frac{a \cdot S}{a + (S - a)e^{-Skt}}$$

mit den Konstanten S (natürliche Schranke), k (Wachstumskonstante), a (Anfangsbestand zur Zeit $t = 0$).

Logistische Funktionen werden z.B. zur Modellierung von Krankheitsausbreitungen verwendet. Als Beispiel kann man eine Schule mit 3000 Schülern betrachten, deren einer am Coronavirus erkrankt ist. Wir nehmen an, dass jeder sich anstecken kann. Die Ausbreitung der Krankheit wird der obigen Funktion folgen und aus früheren Fällen kennt man die Wachstumskonstante k: $P(t) = \frac{3000}{1+2999 \cdot e^{-0.4 \cdot t}}$. D.h. nach 10 Tagen werden sich 54 Personen angesteckt haben.

Logistische Funktion

Wenn man die Krankheitsausbreitung darstellt, sieht man den typischen S-förmigen Verlauf der logistischen Funktion.

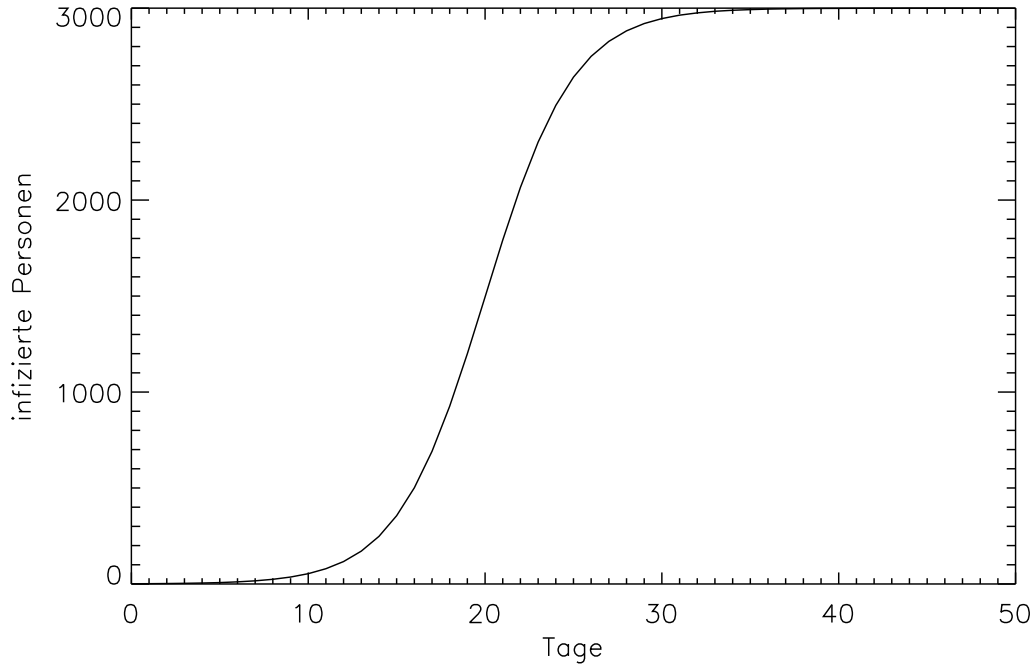


Abbildung: Anzahl infizierter Personen abhängig von der Zeit gemäss vorigem Beispiel.

Die logistische Funktion garantiert, dass sich der Maximalwert 3000 annähert, da irgendwann keine neuen Personen mehr infiziert werden können. Man sieht auch, dass ab einem gewissen Zeitpunkt die Anzahl der Infizierten pro Tag rasant zunimmt und man eine Ausbreitung nur früh stoppen kann.

Logistische Regression

Wir benützen die logistische Funktion $h(z) = \frac{\exp(z)}{1+\exp(z).}$ Mit dieser Funktion kann man die Gleichung für π_i vereinfachen zu

$$\pi_i = h(\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p).$$

Durch Umformen können wir die Gleichung schreiben als

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p$$

wobei $\frac{\pi_i}{1-\pi_i}$ die Chances ("odds") und $\ln \frac{\pi_i}{1-\pi_i}$ die logarithmischen Chancen ("Logit") darstellt. $\ln$ ist der natürliche Logarithmus, der in Fahrmeir's Buch verwirrenderweise als $\log$ geschrieben wird. Die Regressionsparameter können allerdings nicht so einfach bestimmt werden, wie bei der linearen Regression und wir führen dazu die maximum likelihood Methode ein.

Eine Zusammenfassung der logistischen Regression kann hier gefunden werden https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/zusammenhaenge/lreg.html und in Kapitel 12.3

Logistische Regression - Beispiel

Als Beispiel wählen wir einen Fall, wo sowohl die abhängige wie auch die unabhängige Variable kategorial sind.

Beispiel

Eine Fussballmannschaft kann die folgende Statistik an Heim (H) und Auswärtsspielen (A) vorweisen:

	H	A
gewonnen	63	62
verloren	10	24

Bestimmen Sie ein Modell, das den Spielerfolg beschreibt.

Logistische Regression - Beispiel

Als erstes kann man die Wahrscheinlichkeiten berechnet, bei Heimspielen und Auswärtsspielen zu gewinnen.

Gewinn Heimspiel: $\frac{\text{gewonnene Heimspiele}}{\text{total Heimspiele}} = \frac{63}{63+10} = 0.863$ (also 86.3%)

Gewinn A-spiel: $\frac{\text{gewonnene A-spiele}}{\text{total A-spiele}} = \frac{62}{62+24} = 0.721$ (also 72.1%)

Die Chancen, in einem Heimspiel zu gewinnen, sind: $63/10 = 6.3$. Chancen (Odds) können alle Werte ≥ 0 annehmen. Hohe Werte heissen, dass das Ereignis eher eintreten wird. Man kann den identischen Wert auch über die Wahrscheinlichkeiten erhalten: $\frac{\pi_i}{1-\pi_i} = \frac{0.863}{1-0.863} = 6.3$

Logistische Regression - Beispiel

Für eine grafische Darstellung kann man ein Diagramm des folgenden Typs wählen.

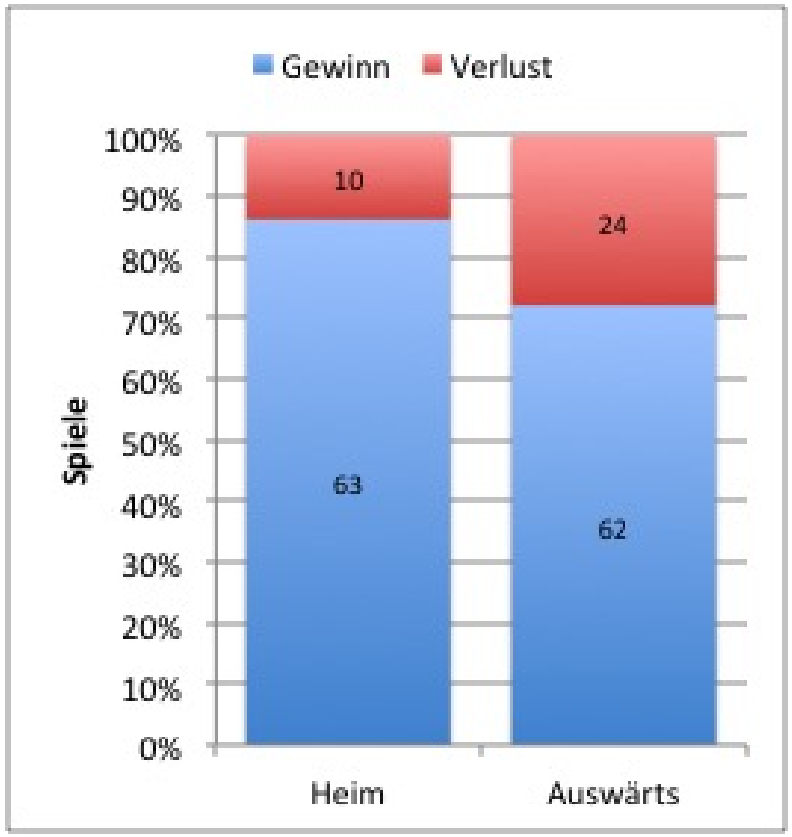


Abbildung: Grafische Darstellung des Beispiels

Logistische Regression - Beispiel

In diesem einfachen Fall kann man das Gleichungssystem schreiben als

$$\ln \frac{\pi_0}{1-\pi_0} = \ln 6.3 = \beta_0 + x_{01}\beta_1 = \beta_0 + 1 * \beta_1$$
$$\ln \frac{\pi_1}{1-\pi_1} = \ln(62/24) = \beta_0 + x_{11}\beta_1 = \beta_0 + 0 * \beta_1$$

Wir haben also Heimspiele mit 1 codiert (Dummyvariable) und Auswärtsspiele mit 0. Durch Ausrechnen gibt dies

$$\beta_0 = 0.949$$
$$\beta_1 = 0.891$$

Man hat somit die Koeffizienten des Modells erhalten. In den meisten Fällen, also wenn man nicht nur 2 Variablen und Messungen hat, benützt man für die Berechnung ein Statistikprogramm (z.B. R).

Logistische Regression - Beispiel

Hier betrachten wir ein weiteres Beispiel, aber mit mehr Datenpunkten: In einer fiktiven Stadt nimmt die Anzahl der Raucher mit der Distanz zum Stadtzentrum zu, wie in der folgenden Grafik dargestellt.

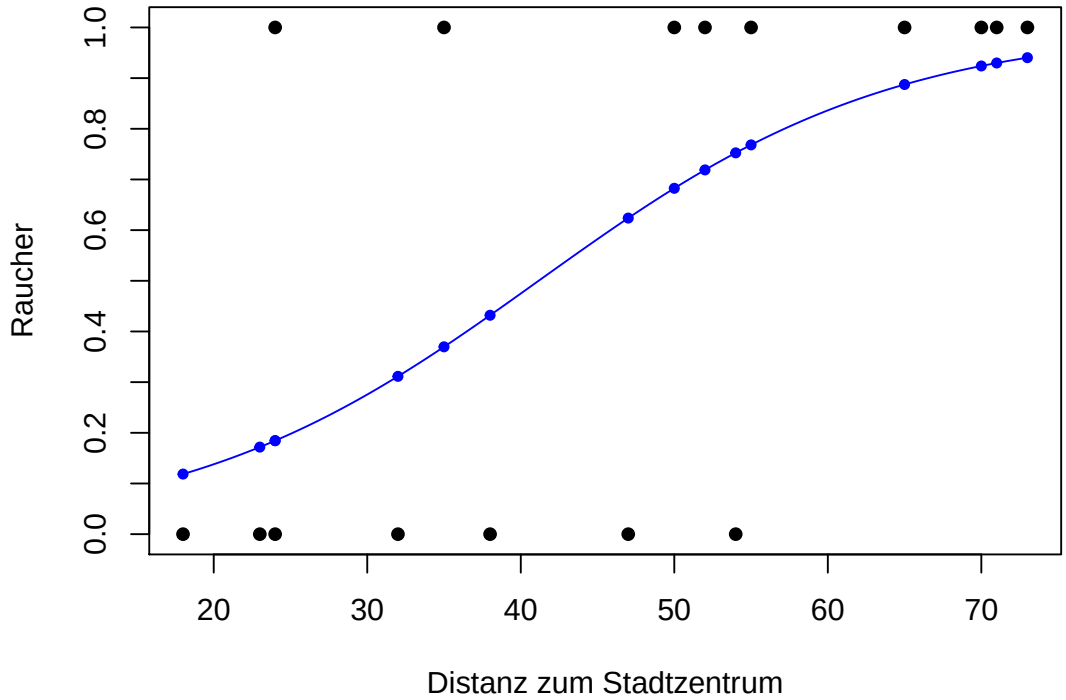


Abbildung: Beispiel mit einer grösseren Anzahl Datenpunkte

In diesem Fall würde das blaue Modell mit dem Computer bestimmt werden, da dies von Hand äusserst mühsam wäre.

Logistische Regression - Beispiel

Das Modell gibt uns die Wahrscheinlichkeit für Raucher (codiert als 1) und Nichtraucher (codiert als 0). Ab einer Distanz von ca. 41 sagt das Modell mit höherer Wahrscheinlichkeit Raucher voraus (Wahrscheinlichkeit> 0.5), bis zu einer Distanz von 41 erwartet man eher Nichtraucher.

Ein Modell ist meistens nicht perfekt und somit gibt es Datenpunkte, die falsch vorausgesagt werden, z.B. würde man den Raucher bei x=24 nicht erwarten.

Wie gut ein Modell zutrifft, kann man anhand einer Klassifikationsmatrix feststellen.

Logistische Regression - Klassifikationsmatrix

Man unterscheidet bei einer Klassifikationsmatrix, ob das Modell eine richtige Vorhersage getroffen hat oder nicht. Unser Modell hier gibt die Wahrscheinlichkeit für einen Raucher. Die richtigen und falschen Klassifikationen lassen sich in die vier folgenden Fälle einteilen:

- Die Person ist Raucher und das Modell hat dies richtig vorhergesagt (richtig positiv)
- Die Person ist Nichtraucher und das Modell hat einen Raucher vorausgesagt (falsch positiv)
- Die Person ist Nichtraucher und das Modell hat dies richtig vorhergesagt (richtig negativ)
- Die Person ist Raucher und das Modell hat einen Nichtraucher vorhergesagt (falsch negativ)

Logistische Regression - Klassifikationsmatrix

In unserem Beispiel kann man dies direkt ablesen, indem man die Datenpunkte im jeweiligen Quadranten zählt.

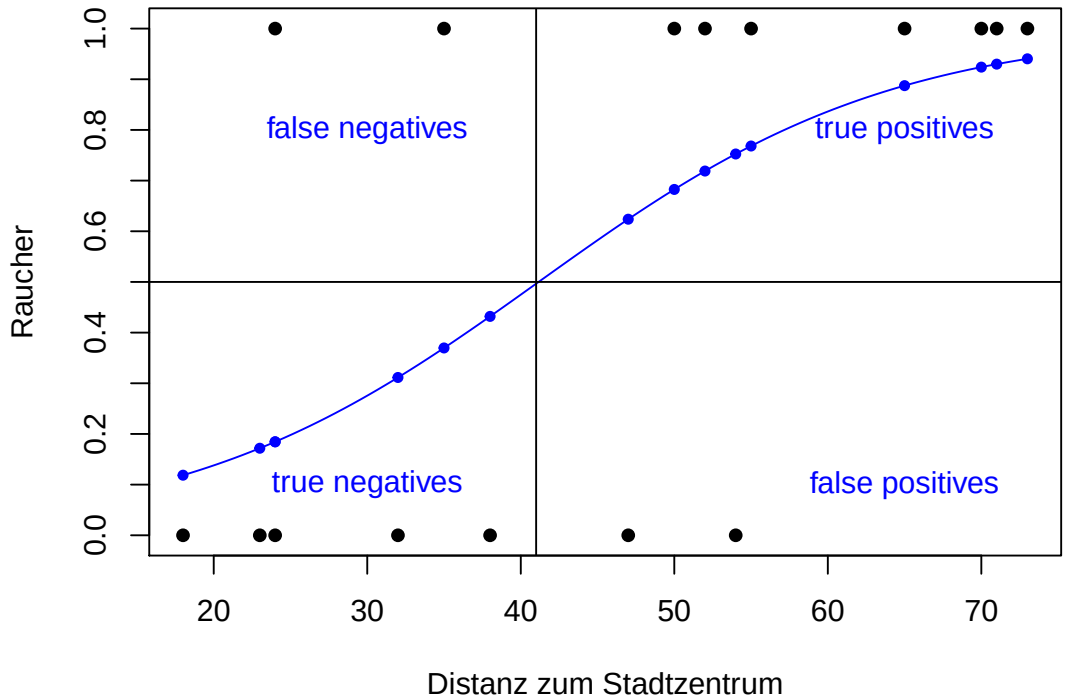


Abbildung: Beispiel mit Hilfslinien und Beschriftungen, um die Klassifikationsmatrix zu bestimmen.

Achtung: die logistische Funktion kann auch bei 1 beginnen und auf 0 gehen (z.B. wenn im Beispiel die Nichtraucher weiter weg von der Stadt wohnen würden). In diesem Fall muss man die Beschriftungen der Quadranten entsprechend anpassen.

Logistische Regression - Klassifikationsmatrix

Üblicherweise schreibt man das Resultat in eine Matrix, die Klassifikationsmatrix genannt wird.

		Status	
		Raucher	Nichtraucher
Vorhersage	Raucher	7 (true positive)	2 (false positive)
	Nichtraucher	2 (false negative)	5 (true negative)

Logistische Regression - Klassifikationsmatrix

Auf English wird die Klassifikationsmatrix oft “confusion matrix” genannt. Sie erlaubt auch, die Sensitivität und Spezifizität zu berechnen, was bei Machine Learning Modellen und deren Validierung äusserst wichtig ist.

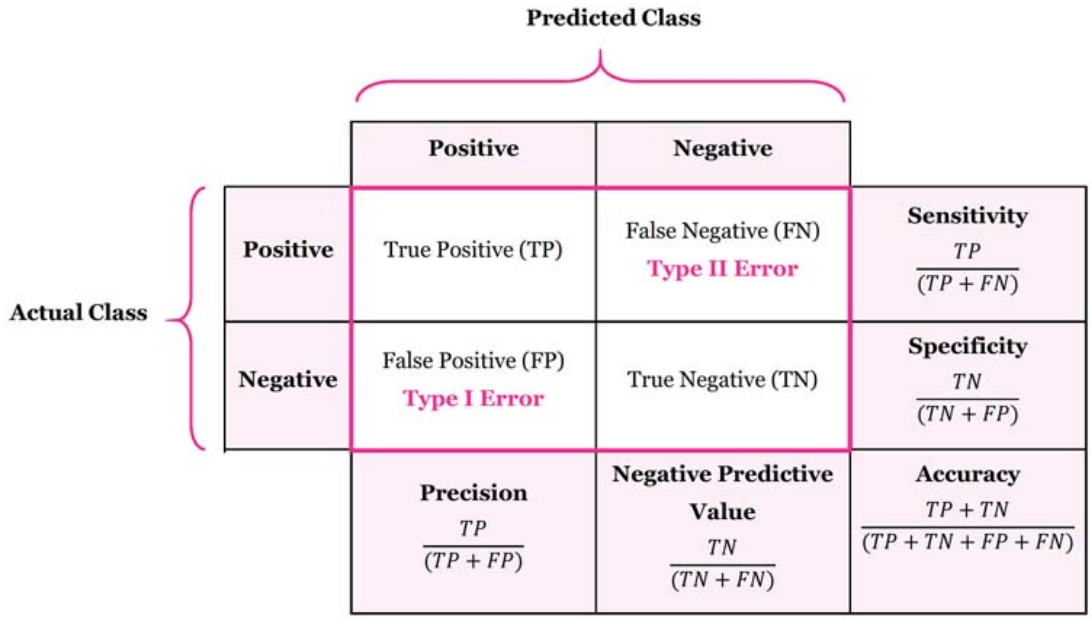


Abbildung: Quelle: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

Beispiel: Bei Covid-Tests werden immer auch Spezifizitäten (z.B. 99.7% Herstellerangabe bei Roche) und Sensitivitäten (z.B. 96.5% bei Roche) angegeben. Die Spezifität von Tests beschreibt die Wahrscheinlichkeit, dass ein negatives Testergebnis auch wirklich negativ ist, was hier mit 99.7% sehr gut ist. Nur 0.3% der Tests geben somit ein positives Ergebnis, obwohl die Person kein Covid hat (false positive). Die Sensitivität von Covid-Tests beschreibt die Wahrscheinlichkeit eines positiven Testergebnisses, welches auch richtig als positiv erkannt wurde. Wenn man also Covid hat und einen Roche-Schnelltest macht, dann zeigt er in 96.5% der Fälle positiv an. Dies heisst aber, dass man in 3.5% der Fälle ein negatives Ergebnis erhält, aber eigentlich infiziert ist (false negative).

Maximum Likelihood Methode

Wenn man eine sehr grosse Grundgesamtheit hat, ist es zu aufwändig, alle einzelnen Daten zu analysieren und deswegen nimmt man eine Stichprobe. Die Maximum Likelihood Methode ist ein parametrisches Schätzverfahren, mit welchem man aus der Stichprobe die Parameter der Grundgesamtheit (z.B. Erwartungswert oder Standardabweichung) schätzt.

Die Idee ist, dass man anhand der Stichprobe die Werte der wahren Parameter so schätzt, dass die beobachtete Stichprobe die wahrscheinlichste Verteilung widerspiegelt.

Maximum Likelihood Methode

Beispiel

Ein gut erklärtes Beispiel ist in diesem Video:
https://www.youtube.com/watch?v=XepXtl9YKwc

Maximum Likelihood Methode

Beispiel

400 Leute kaufen täglich im Dorfladen ein. Man will wissen, wie viele Prozent ihre Einkaufstasche selber mitbringen. Dazu befragt man eine Stichprobe von 30 Kunden und 12 sagen, sie haben eine Tasche dabei, 18 nicht.

Erinnerung: Die Binomialverteilung ist eine der wichtigsten diskreten Wahrscheinlichkeitsverteilungen und definiert als $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$. Wir möchten also die Wahrscheinlichkeit der Binomialverteilung bestimmen, genau 12 von 30 Kunden auszuwählen, in Abhängigkeit des unbekannten Anteilswerts p der Grundgesamtheit.

Wenn $p = 30\%$ aller Kunden Einkaufstaschen dabei hätten (n=30, k=12, p=.3), wäre $P = 7.49\%$. Bei $p = 40\%$ aller Kunden wäre $P = 14.7\%$, bei $p = 50\%$ aller Kunden wäre $P = 8.06\%$. Man kann berechnen, dass die maximale Wahrscheinlichkeit bei der Annahme p=40% erreicht wird und somit ist dies der Maximum Likelihood Schätzer und die wahrscheinlichste Annahme für die Grundgesamtheit. Diese Lösung ist logisch, weil auch 12/30 40% entsprechen.

Maximum Likelihood Methode

Allgemein schreibt man bei der Maximum Likelihood Methode die Wahrscheinlichkeitsdichte f abhängig von einem Parameter θ. Wir haben eine Zufallsstichprobe mit n Realisierungen $x_1, ..., x_n$.

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Statt dass man θ konstant annimmt und die Dichte für beliebige Werte $x_1, ..., x_n$ auswertet, kann man umgekehrt für feste Realisationen $x_1, ..., x_n$ die gemeinsame Dichte als Funktion von θ betrachten. Dies ergibt die Likelihood Funktion

$$L(\theta) = f(x_1, ..., x_n|\theta)$$

Das Maximum Likelihood Prinzip maximiert diese Funktion.

Maximum Likelihood Prinzip

Satz (Maximum Likelihood Prinzip)

Das Maximum Likelihood Prinzip besagt: Wähle zu $x_1, ..., x_n$ als Parameterschätzung denjenigen Parameter $\hat{\theta}$, für den die Likelihood maximal ist, d.h.

$$L(\hat{\theta}) = \max_{\theta} L(\theta)$$

Üblicherweise bestimmt man das Maximum einer Funktion durch Ableiten und Nullsetzen der Ableitung. Für die Likelihood ist das wegen der Produkte in $L(\theta)$ meist mühsam. Statt der Likelihood selbst kann man stattdessen die logarithmierte Likelihood, die sogenannte Log-Likelihood, maximieren. Da Logarithmieren eine streng monoton wachsende Transformation ist, liefert das Maximieren von $L(\theta)$ und $\log L(\theta)$ denselben Wert $\hat{\theta}$. Für den bisher betrachteten Fall unabhängiger und identischer Wiederholungen ergibt sich die Log-Likelihood als Summe

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Maximum Likelihood Beispiele

Beispiele zur Berechnung der Maximum Likelihood finden sich z.B. hier:
https://de.wikipedia.org/wiki/Maximum-Likelihood-Methode
und in Kapitel 9.3

Ausserdem wird im JiTT ein Beispiel vorgerechnet.

Section 3

Liste der Änderungen

Liste der Änderungen

- 25.9.20: Update Lernmaterialien Links. S.18 α und β waren vertauscht. Neue Folie S.35 (log. Funktion).
- 21.3.21: Update R^2: Detailliertere Erklärung und Formel.
- 11.6.21: Klassifikationsmatrix neu.
- 30.9.21: Klassifikationsmatrix erweitert.