Regression

Korrelation
simple lineare Regression
kurvilineare Regression
Bestimmtheitsmaß und Konfidenzintervall

Zusammenhänge zw. Variablen

Betrachtet man mehr als eine Variable, so besteht immer auch die Frage nach den Abhängigkeiten und Zusammenhängen zwischen diesen Variablen.

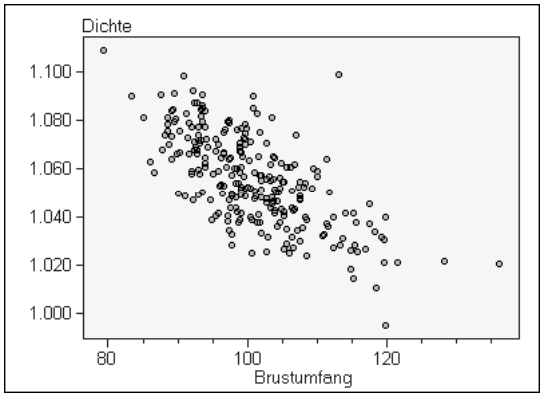
Einfachster Ausdruck von Zusammenhängen: Streuplots

Zusammenhang:

qualitativ
(“hängt ab von...”)

quantitativ
(Maßzahl für den Zusammenhang)

Modellierung
(mathemat. Funktion y = f(x))



Korrelation 1

Problem der Kovarianz: keine Invarianz gegen Skalierung

Lösung: Standardabweichung als Maß verwenden

---> Korrelationskoeffizient nach Pearson

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Eigenschaften:

Werte zwischen -1.0 und +1.0

Annahmen:

- linearer Zusammenhang zwischen x und y
- kontinuierliche Zufallsvariablen
- beide Variablen müssen normal verteilt sein
- die Streuungen von x und y müssen voneinander unabhängig sein

Korrelation 2

Bestimmtheitsmaß:

Quadrat des Korrelationskoeffizienten

wird zur Gütebeurteilung von linearen Modellen verwendet

Tipp für die Praxis:

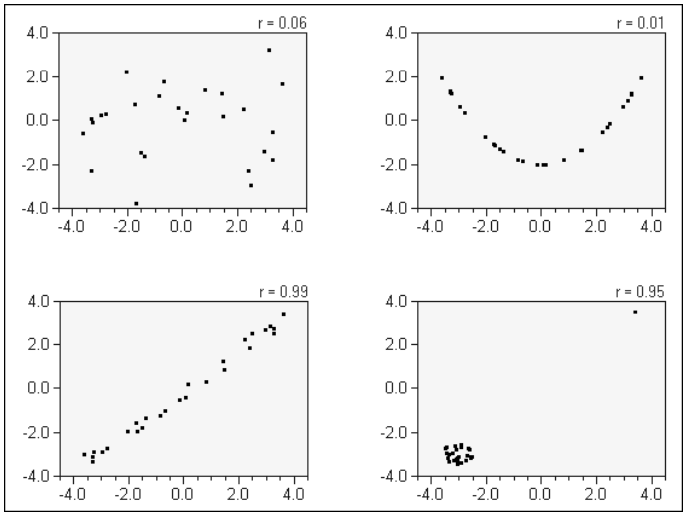
Die Standardformel des Korrelationskoeffizenten ist umständlich und rechenaufwändig. Für den Einsatz im Computer eignet sich eine andere Formel besser (siehe Teach/Me)

Übung:

Schätzen des Korrelationskoeffizienten

[GrundStat: PEARCORR] Direktlink: Pearsons Korrelationskoeffizient

Korrelation 3



Korrelation =

- Maß für lineare Zusammenhänge
- empfindlich auf Ausreißer

Beispiele:

[GrundStat: CORREX]
Direktlink: Beispiele von Korrelationen

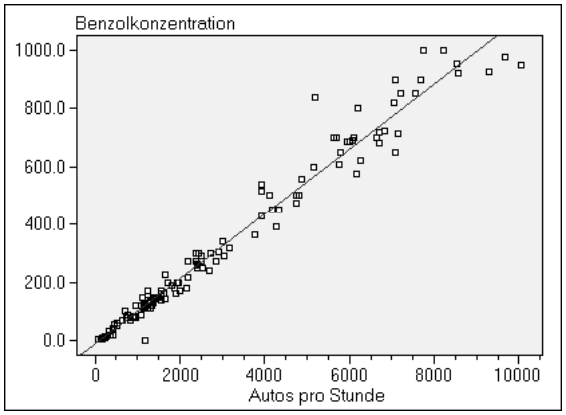
Regression 1

Linearer Zusammenhang zwischen unabhängiger und abhängiger Variable

==> einfachstes Modell ist eine Gerade

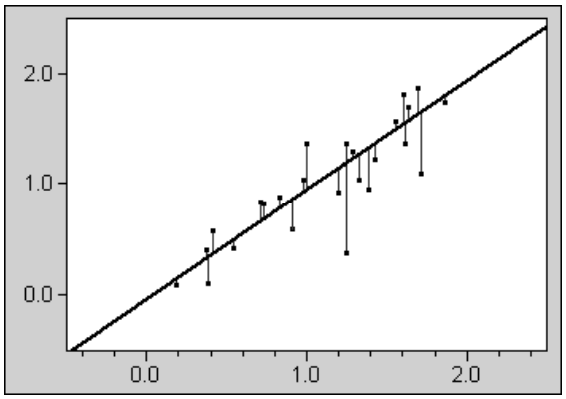
Parameter der Geraden
y = k*x + d
so bestimmen, dass die Gerade den Zusammenhang "optimal" wiedergibt.

Was heißt "optimal"?



Regression 2

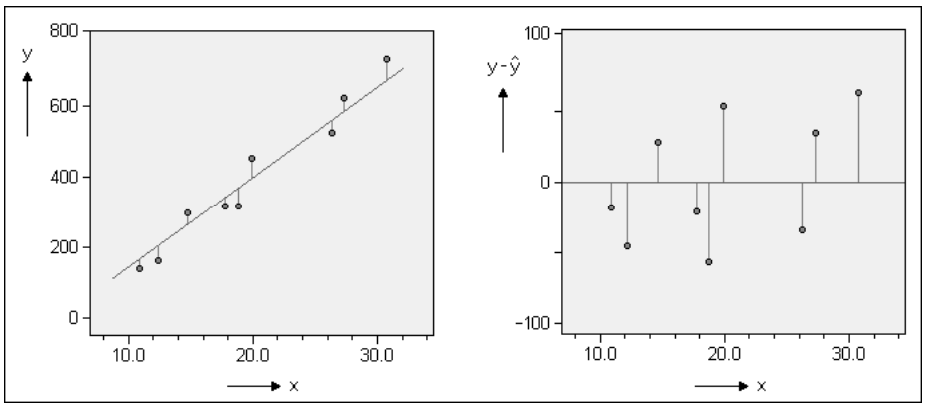
Regression kann als Optimierungsproblem aufgefasst werden: Gerade so legen, dass Summe der Abstände (rote Linien) minimal wird.



[GrundStat: REGOPTI]
Direktlink: Regression - Gerade

Residuen

Residuen (sing. *Residuum*) sind die "Restwerte", also die Differenzen der y-Werte zwischen Regressionsgerade und den Messwerten:



Annahmen und Voraussetzungen

Der erwartete Zusammenhang zwischen X und Y ist linear

Unterscheidung zwischen linearen, krummlinigen (kurvilinearen) und nicht linearen Zusammenhängen.

Alle Messungen sind voneinander unabhängig

Jeder Trend über die Zeit oder eine gemeinsame Korrelation mit einer dritten Variablen müssen vermieden werden.

Für jedes X sind die Y-Werte normal verteilt.

In der Praxis oft schwer zu überprüfen (geringe Zahl an Messwerten)

Für jedes X hat die Y-Verteilung dieselbe Varianz

Regression arbeitet nur mit homoskedastischen Daten korrekt

Analyse der Residuen

Die Analyse der Residuen ist ein einfaches und effizientes Mittel die Regression zu überprüfen:

Residuen müssen

- um Null symmetrisch verteilt sein
- eine Normalverteilung bilden
- unabhängig vom Wert der unabhängigen Variable (meist x) sein

DirektLink: Analyse der Residuen

Skedastizität

Skedastizität gibt das Verhalten der Streuung in Abhängigkeit von der unabhängigen Variablen an. Zwei Fälle:

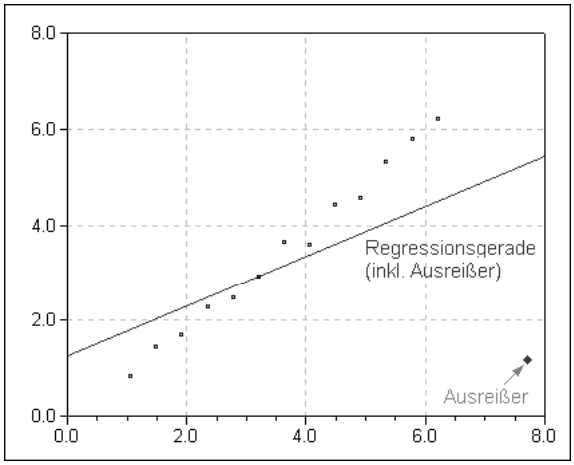
homoskedastisch: die Streuung ist konstant

heteroskedastisch: die Streuung ändert sich mit der Signalamplitude

Skedastizität kann anhand des Residuen-Plots erkannt werden

[GrundStat: SCEDAST] Direktlink: Skedastizität

Hebeleffekt



Ausreißer haben großen Einfluss auf Regression:

ein falscher Wert kann die Regressionsgerade massiv stören! --> daher immer visuelle Kontrolle der Regression notwendig

[LEVERAGE]
Direktlink: Hebeleffekt

Zuverlässigkeit - 1

Zuverlässigkeit von Regressionsmodellen:

hängt ab von den Residuen, der Zahl der Messwerte und (bei multivariaten Modellen) von der Zahl der Variablen.

Bei simpler linearer Regression wird meist das Bestimmtheitsmaß (engl. *coefficient of determination*, auch *goodness of fit*) herangezogen.

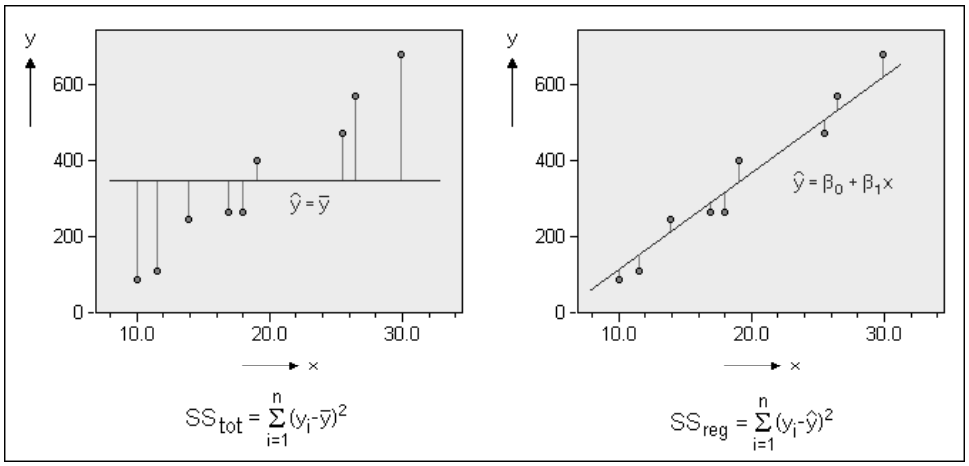
Idee: Die Verkleinerung des Modellfehlers (=Residuen) bei Einbeziehung der unabhängigen Variable (siehe nächste Seite)

Bestimmtheitsmaß r^2

Bei simpler linearer Regression ist r^2 gleich dem Quadrat des Korrealtions-koeffizienten zwischen unabhängiger und abhängiger Variable.

Bereich: $0.0 <= r^2 <= 1.0$

Zuverlässigkeit - 2



$$r^2 = \frac{SS_{tot} - SS_{reg}}{SS_{tot}}$$

Zuverlässigkeit - 3

Ist die Steigung k der Regressionsgeraden ungleich null?

Die Antwort auf diese Frage gibt ebenfalls einen Hinweis auf die Zuverlässigkeit eines Regressionsmodells, da für den Fall, dass k sich *signifikant* von null unterscheidet die gefundene Regressionsgleichung kein Zufallsergebnis mehr sein kann.

Zur Überprüfung verwendet man den F-Test aus der ANOVA (analysis of variances). Übersteigt der F-Wert die kritische Grenze, so wird die Nullhypothese k=0 verworfen (das Modell ist also gültig).

Ursprung der Variationen	Freiheits-grade	Summe der Quadrate	Mittelwert der Quadratsumme	F-Wert
Regression	1	$SS_{reg} = \sum (\hat{Y}_i - \bar{Y})^2$	$MS_{reg} = SS_{reg}$	MS_{reg}/MS_{res}
Residuen	n-2	$SS_{res} = \sum (Y_i - \hat{Y}_i)^2$	$MS_{res} = SS_{res} / n-2$	
Gesamt	n-1	$SS_{tot} = \sum (Y_i - \bar{Y})^2$		

n = Zahl der Messwerte

Zuverlässigkeit - 4

Szenario:

Bestimmtheitsmaß nahe 1.0
F-Wert weit über 1000

Kann eine solche Regression "falsch" sein ?

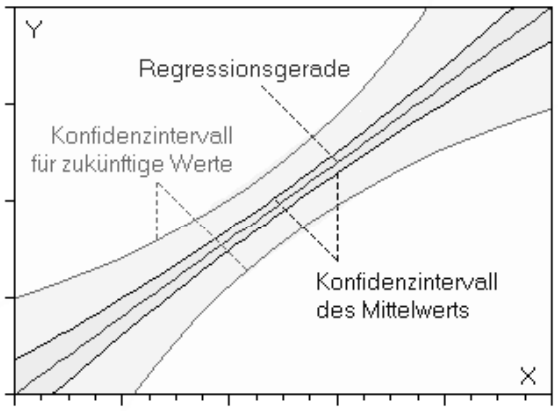
Ja! Wenn die Anforderung an die Präzision der Kalibration sehr hoch sind.

Klassisches Beispiel: Eichung eines Massenspektrometers
hier sind die Anforderungen an die Präzision so hoch, dass auch bei unbrauchbarer Kalibration sich ein Bestimmtheitsmaß von > 0.9999 ergibt

Grund: die Abweichungen von der Modellkurve sind zwar klein im Vergleich zum Kalibrationsbereich, aber (zu) groß im Vergleich zu den Anforderungen an die Präzision und Genauigkeit ---> Residuenplot ansehen!!

Konfidenzintervall

gibt den Vertrauensbereich der Regression an



2 Intervalle:

1. Konfidenzintervall des Mittelwerts

2. Konfidenzintervall für geschätzte Werte

Das zweite Intervall ist für die Praxis wichtig und ist immer deutlich größer als das erste.