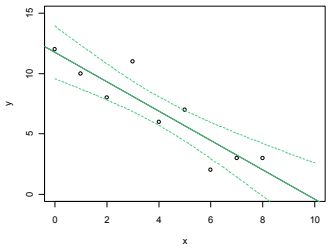
HOCHSCHULE
WEIHENSTEPHAN-TRIESDORF
UNIVERSITY OF APPLIED SCIENCES

STATISTISCHE AUSWERTUNG ÖKOLOGISCHER DATENSÄTZE

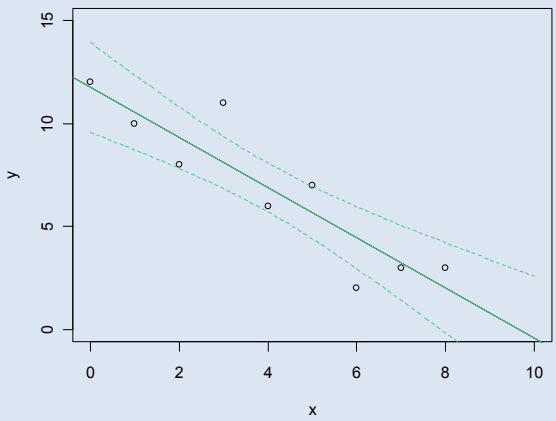
REGRESSION

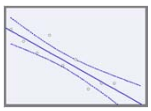


Studiengänge US, TE, WT, 3. & 7. Semester

Prof. Dr. Michael Rudner

Einfache lineare Regression





Lineare Regression

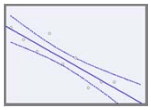
Beziehung zwischen einer abhängigen Variablen (AV)
und einer oder mehreren unabhängigen Variablen (UV).

- Eine abhängige Variable
- Eine unabhängige (erklärende) Variable:
  einfache (univariate) Regression
- Mehrere unabhängige Variablen:
  multiple Regression

• Kausal
• Quantitativ — je...desto...

$Y = f(X)$

$Y = f(X_1, X_2 \ldots, X_j, \ldots X_J)$

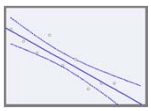


Lineare Regression

Voraussetzungen:

• Alle Variablen sollten kontinuierlich sein.
• Ggf. Umformung von kategorialen Variablen
  in Dummy-Variablen (0,1)
• Wenn AV nominal (z.B. Anwesenheit: ja, nein)
  dann logistische Regression
• Lineare Regression:
  linearer Zusammenhang zwischen AV & UV

$$y = a + b \cdot x \quad \textit{oder} \quad y = b_0 + b_1 \cdot x$$



Lineare Regression

Vorteile:
- sehr flexibel
- Prognose für abhängige Variable

Typische Fragestellungen:

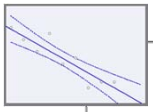
Ursachenanalysen:

Welche Faktoren beeinflussen AV?

Wie verändert sich AV (abhängige Variable) entlang eines Gradienten der UV (unabhängigen Variablen) ?

Wirkungsprognosen

Wenn sich A um Wert X ändert,
welchen Einfluss hat dies auf B?



Regression

$$\hat{Y} = b_0 + b_1 X$$

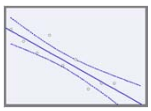
mit

$\hat{Y}$	=	Schätzung der abhängigen Variablen Y	
b_0	=	konstantes Glied	(Achsenabschnitt, *Intercept*)
b_1	=	Regressionskoeffizient	(*Coefficient*)
X	=	unabhängige Variable	

b0 Ursprung der Regressionsgerade
$\hat{Y}$ für X=0

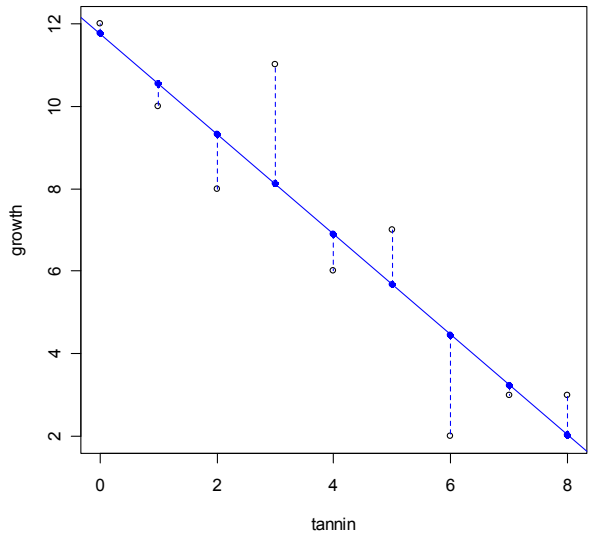
b1 Steigung Regressionsgerade, d.h. um wie viel ändert sich Y, wenn X sich um 1 Einheit ändert

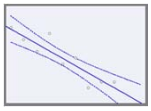
Wichtig zur Wirkungs-abschätzung



Lineare Regression

Residuum:
Abstand des beobachteten Wertes zum Wert, den das Regressionsmodell annimmt.





Lineare Regression

- Residuen (e_k): Abweichungen von Regressionsgerade
- Methode der kleinsten Abstandsquadrate (*Ordinary least squares*)
- Ziel: Summe der quadrierten Residuen soll minimal sein, danach b_0 und b_1 wählen
- Quadrieren, damit sich nicht negative und positive Abweichungen eliminieren

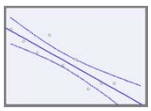
$$e_k = y_k - \hat{y}_k \qquad (k=1, 2, ..., K) \qquad (4)$$

mit

y_k = Beobachtungswert der abhängigen Variablen Y für x_k
$\hat{y}_k$ = ermittelter Schätzwert von Y für x_k
e_k = Abweichung des Schätzwertes von Beobachtungswert
K = Zahl der Beobachtungen

Durch Umformung von (4) und unter Einbeziehung von (2) läßt sich folgende Funktion bilden:

$$Y = \hat{Y} + e$$
$$= b_0 + b_1 X + e \qquad (5)$$



Lineare Regression

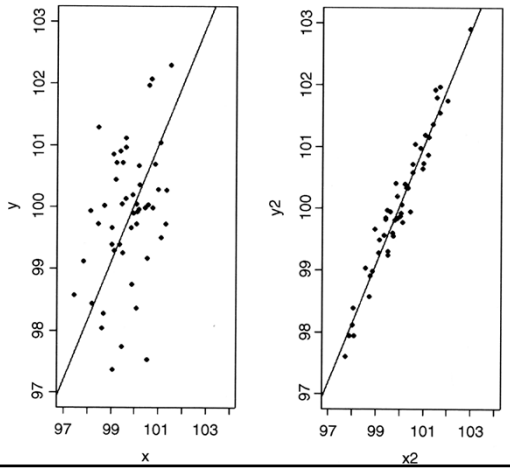
Prüfen der Güte des Modells

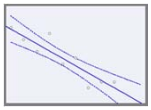
Güte der Regressionsfunktion

F-Statistik

r^2

Standardfehler





Lineare Regression

Prüfen der Güte des Modells

Prüfung der unabhängigen Variablen

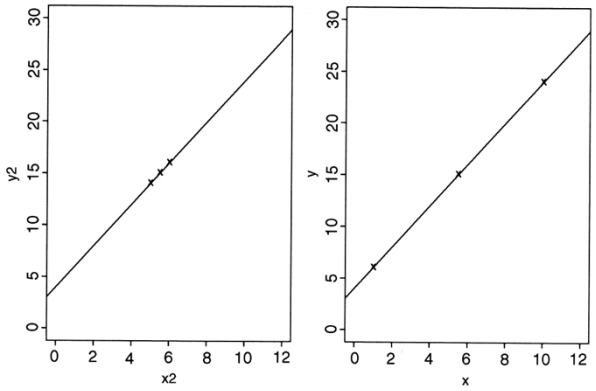
Standardfehler der Steigung

$$s.e._b = \sqrt{\frac{s^2}{SSX}}$$

$$SSX = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

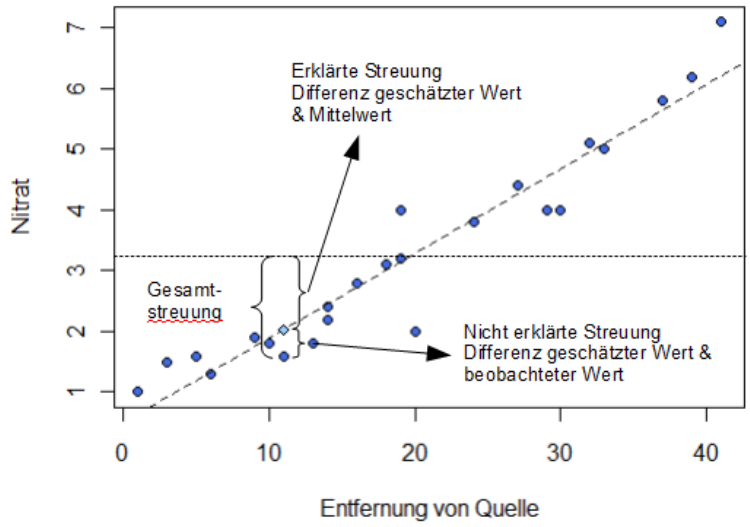
beta-Wert

t-Wert



Lineare Regression

Bestimmtheitsmaß (Determinationskoeffizient)



Lineare Regression

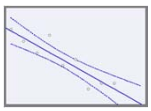
Bestimmtheitsmaß (Determinationskoeffizient)

Basiert auf Schätzung der Residuen für einzelne Beobachtungspunkte

r^2 zwischen 0-1.

In Ökologie in signifikanten Modellen 0.1-0.8

$$R^2 = \frac{\sum_{k=1}^{K}(\hat{y}_k - \bar{y})^2}{\sum_{k=1}^{K}(y_k - \bar{y})^2} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}$$

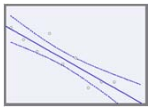


Lineare Regression

Bestimmtheitsmaß (Determinationskoeffizient)

- r^2 = Abschätzung des Effekts, wie viel der Streuung wird durch UV erklärt
- r^2 = Quadrat des Korrelationskoeffizienten r zwischen beobachteten und geschätzten Y-Werten
- r^2 wird von Anzahl der UVs beeinflusst, daher bei vielen UVs häufig korrigiertes r^2 benutzt, dies nimmt mit Anzahl der UVs ab, da Erklärungsanteil jeder UV zufällig bedingt sein kann

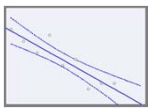
Maß wie gut geschätzte Werte beobachtete Werte abbilden



Lineare Regression

F-Statistik

- Besitzt **Modell** Vorhersagequalität für Grundgesamtheit?
- Prüfung gegen Nullhypothese, dass UV keinen Einfluss auf AV hat
- Wenn Nullhypothese stimmt: F-Wert = 0
  Daher ist F-Statistik ein Test, ob F-Wert unterschiedlich von 0



Lineare Regression

F-Statistik

$$F_{emp} = \frac{R^2 / J}{(1-R^2)/(K-J-1)}$$

- J = Anzahl UV
- K = Stichprobengröße
- Vergleich beobachteter mit theoretischem F-Wert
- Vertrauenswahrscheinlichkeit (0.95 oder 0.99)
- Irrtumswahrscheinlichkeit (5%, 1%)
  α= 1- Vertrauenswahrscheinlichkeit

- oder: zugehörigen p-Wert ermitteln

Lineare Regression - Output

```
> tann<-read.table("Crawley\\tannin.txt",header=T)
> x<-tann$tannin
> y<-tann$growth
> tann.lr <-lm(y~x)
> summary(tann.lr)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4556 -0.8889 -0.2389  0.9778  2.8944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7556     1.0408  11.295 9.54e-06 ***
x            -1.2167     0.2186  -5.565 0.000846 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-Squared: 0.8157,     Adjusted R-squared: 0.7893
F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.000846
```

Verteilung der Residuen

Regressionskoeffizienten

Fehler der Regressionskoeffizienten

p-Wert

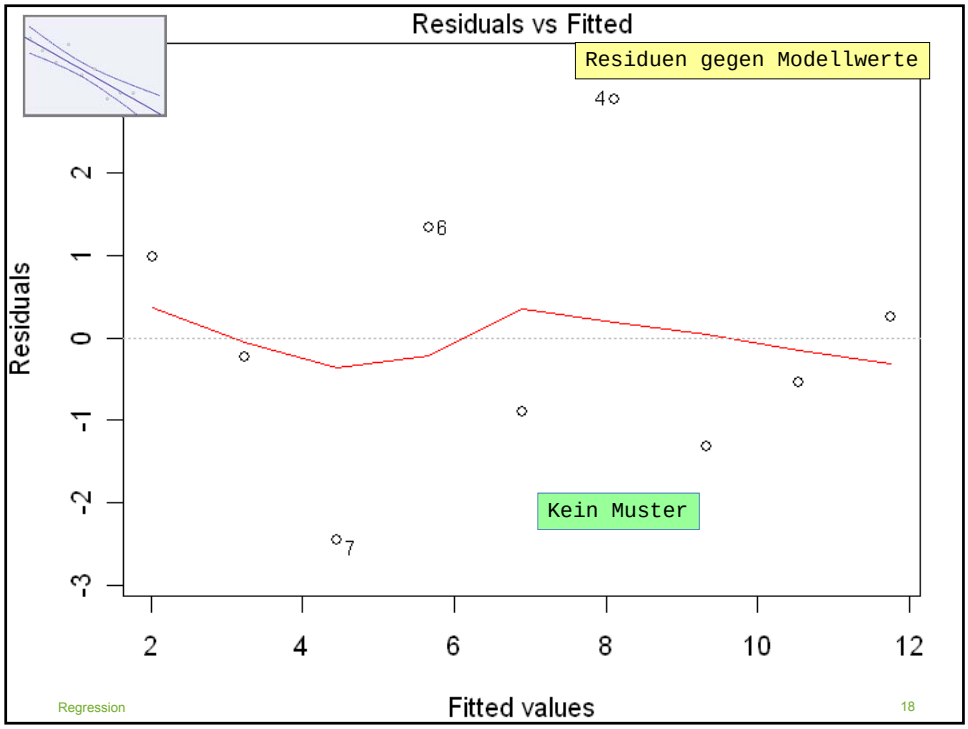
Lineare Regression - Output

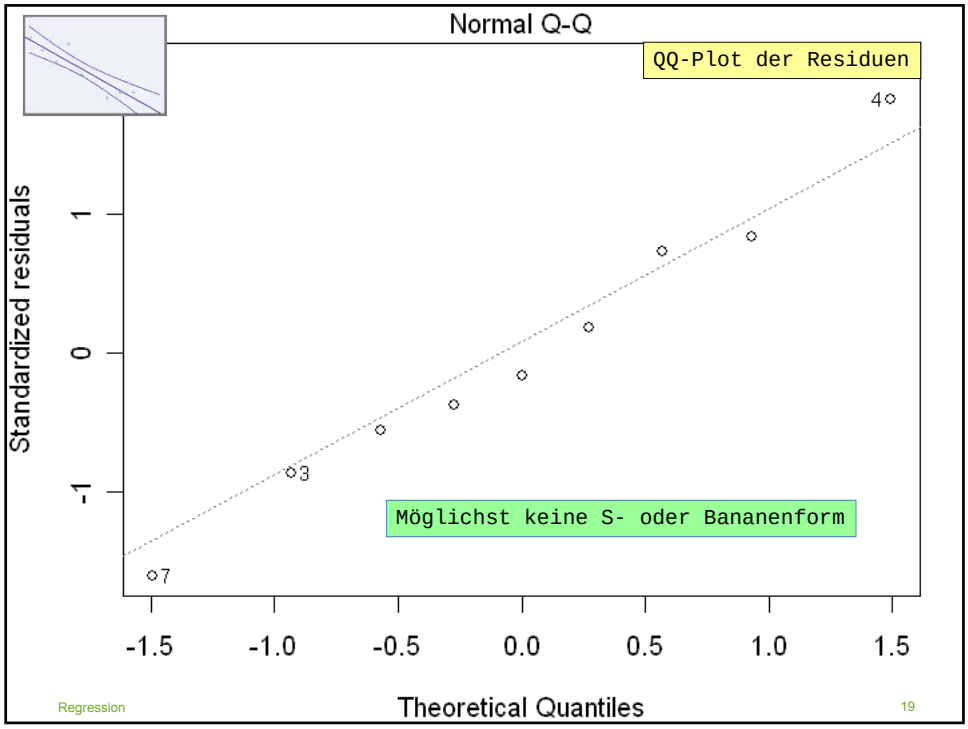
```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4556 -0.8889 -0.2389  0.9778  2.8944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7556     1.0408  11.295 9.54e-06 ***
x            -1.2167     0.2186  -5.565 0.000846 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-Squared: 0.8157,     Adjusted R-squared: 0.7893
F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.000846
```

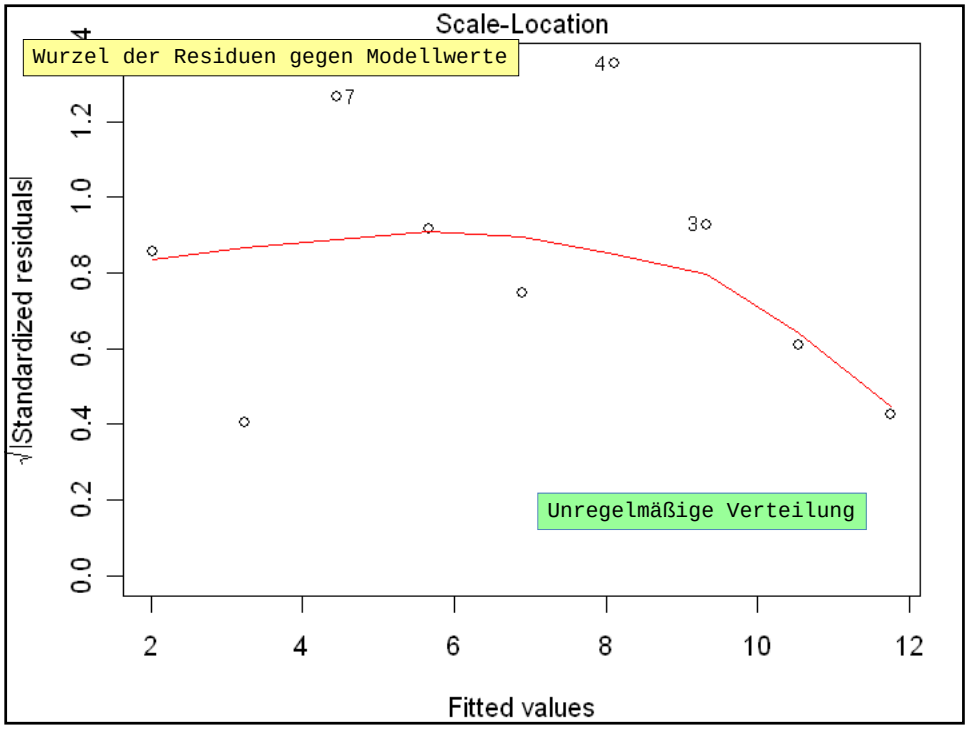
Determinationskoeffizient

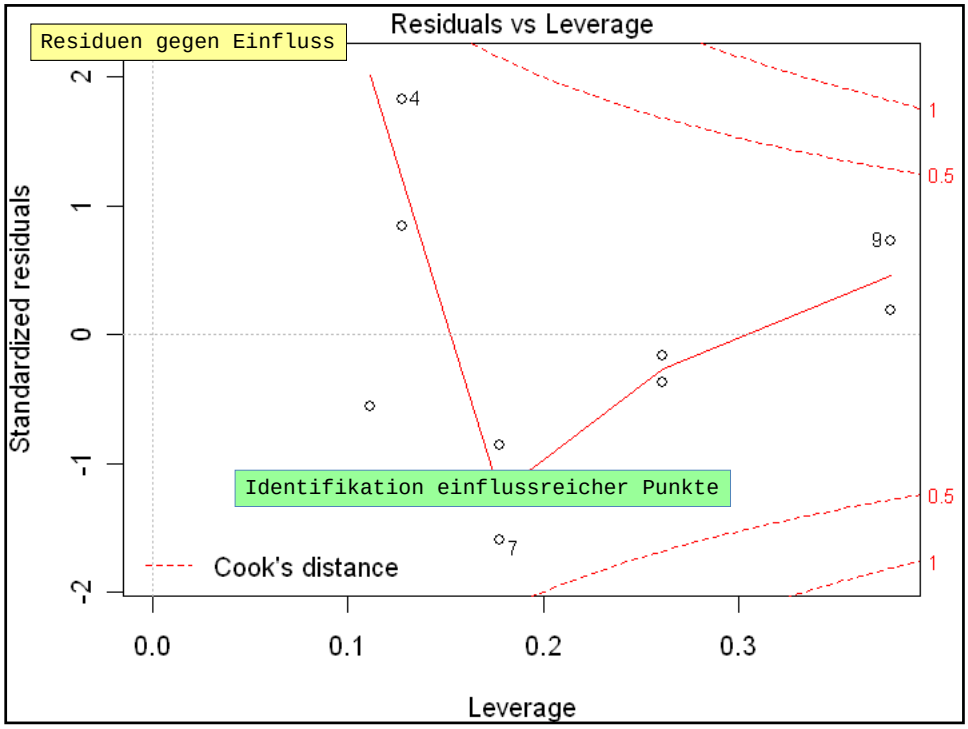
F-Statistik und zugehöriger p-Wert

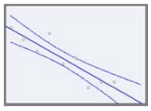
Standardschätzfehler




Normal Q-Q
QQ-Plot der Residuen
Standardized residuals
Theoretical Quantiles
Möglichst keine S- oder Bananenform
Regression
19


Scale-Location
Wurzel der Residuen gegen Modellwerte
√|Standardized residuals|
Fitted values
Unregelmäßige Verteilung





Lineare Regression

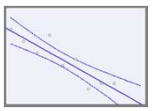
Standard-Schätzfehler

$$s.e._{\hat{y}} = \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}\right]} = s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}}$$

Für Konfidenzintervall der Regressionsgeraden:

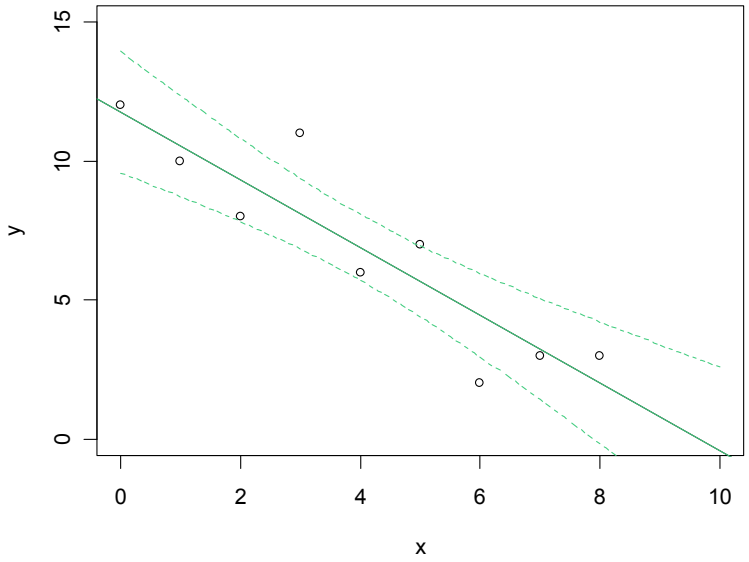
$$\Delta_{krit} = \hat{y}_j \pm t_{(1-\alpha/2)} \cdot s.e._{\hat{y}}$$

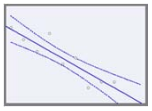
$$\Delta_{krit} = \hat{y}_j \pm t_{(1-\alpha/2)} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}}$$



Lineare Regression

Konfidenzintervall





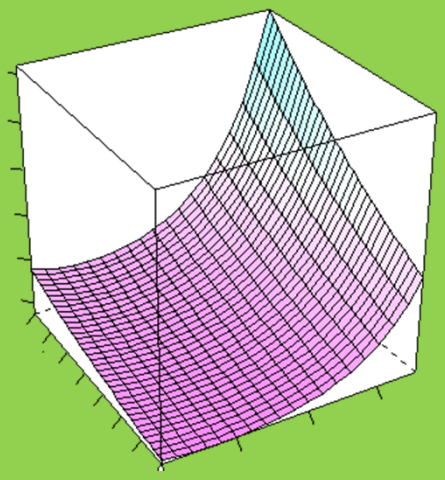
Lineare Regression

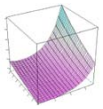
Zusammenfassung

- Regressions-Parameter werden aus Daten geschätzt
- Vielzahl möglicher Modelle für einen Datensatz
- Modellauswahl ist ein wesentlicher Schritt
- Diagnostische Plots um zu prüfen, ob das gewählte Modell passt
- Hauptprobleme:
  - **nicht-konstante Varianz** und
  - **nicht-normalverteilte Fehler**
- Mögliche Lösungen für gekrümmte Zusammenhänge zwischen y und x:
  - Quadrierter Term
  - Transformation

- Vergessen Sie nicht, die Modellwerte rückzutransformieren, bevor Sie die Regressionskurve zeichnen!

Multiple lineare Regression



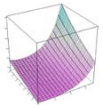


Multiple lineare Regression

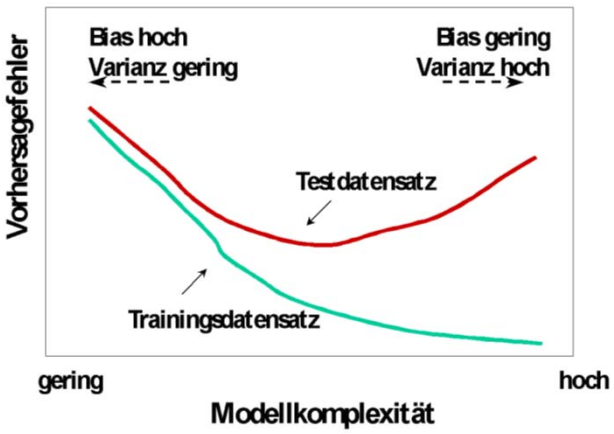
- Eine abhängige Variable
- Mehrere unabhängige Variablen

Wesentliche Fragen:
- Linearer Zusammenhang
- Interaktion der Variablen
- Korrelationen der Variablen
- Überparametrisierung

?

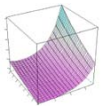


Überparametrisierung



➔ maximal n/3 Modellparameter

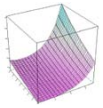
Reineking & Schröder 2004 UFZ-Berichte 9/2004



Multiple lineare Regression

Vorgehen

- Überlegen, welche Variablen sinnvollerweise ins Modell genommen werden sollen
- Prüfen der Korrelationen der Variablen untereinander – Problem der Multikollinearität
  Stichwort: *Variance Inflation Factor*
- Beschränken der Anzahl der Modellparameter auf maximal n/3



Multiple lineare Regression

Korrelation der Variablen

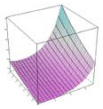
		s	*t*	*v*	*q*
AV	s	1,00			
UV	t	-0,29	1,00		
	v	0,68	0,06	1,00	
	q	0,05	0,61	0,41	1,00

Starke Korrelation

Man sollte stark korrelierte UV nicht gemeinsam in ein Modell aufnehmen.

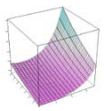
Herausnehmen der Variable „Entfernung zur Quelle“

AV: Sauerstoffgehalt s;
UV: Wassertemperatur t, Fließgeschwindigkeit v, Abstand von der Quelle q



Modellbildung

- Ziel: Kleinstmögliches Modell (*Ockham's razor*)
- Volles Modell aufstellen
- Schrittweise rückwärts unbedeutende Variablen entfernen.
- Hinweis aus t-Statistik
- Minimales zufriedenstellendes Modell weiter prüfen
- ggf. Transformationen der Werte
- ggf. entfernen zu einflussstarker Werte
- Modellvergleich mit F-Test



Responseoberfläche

- Erzeugen einer Oberfläche der Modellwerte in Abhängigkeit von zwei erklärenden Variablen
- weitere Prädiktoren müssen berücksichtigt werden, da sie Einfluss auf den Achsenabschnitt haben
- also zwei Variablen zur Darstellung auswählen, und für die restlichen Variablen Referenzwerte auswählen, die anzugeben sind.
- ggf. mehrere Grafiken mit unterschiedlichen Referenzwerten erzeugen