



# **Skript: Einführung in die Statistik**

## **Grundlagen der Mathematik II Lineare Algebra und Statistik**

**FS 2010**

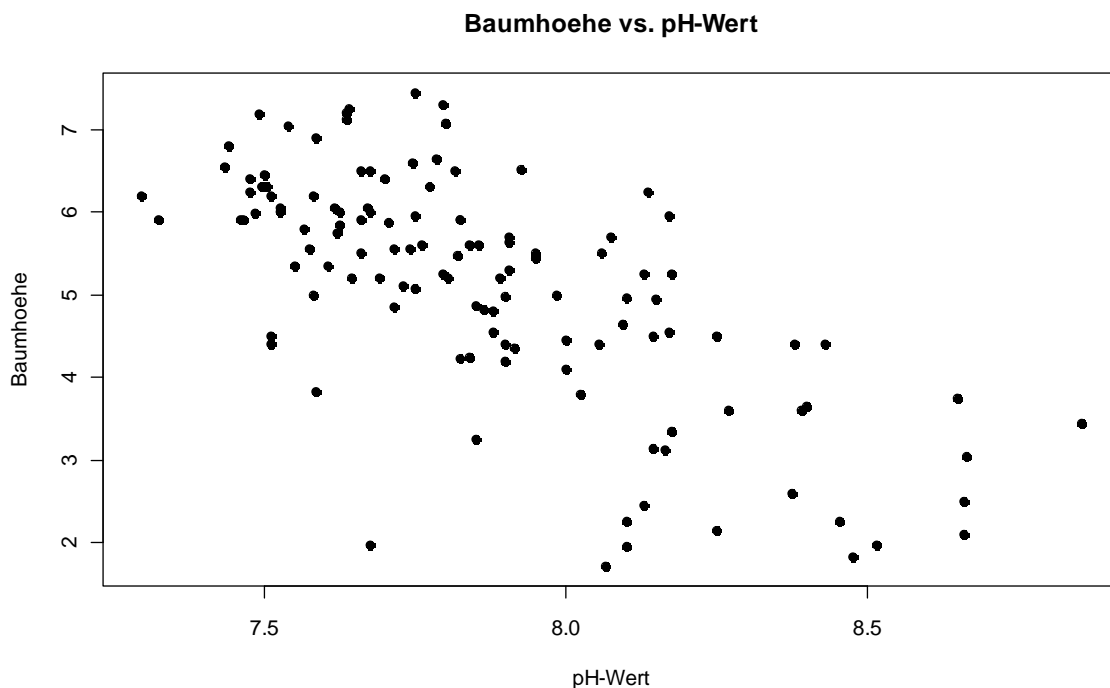
Dr. Marcel Dettling  
provisorische Version Kapitel Regression  
2. Juni 2010

# 1 Einfache lineare Regression

## 1.1 Einführung, Beispiel, Fragestellung

In der Wissenschaft, wie auch im Alltag fragen wir uns immer wieder, wie eine Grösse von Interesse von anderen Einflussfaktoren abhängt. Diese grundlegende Frage wird durch die *Regression* quantitativ untersucht. So erstaunt es auch nicht, dass es sich bei der Regression um eine der am meisten angewandten Techniken der Statistik handelt. Wir betrachten das folgende Beispiel:

**Beispiel:** In Indien behindern basische Böden Pflanzen beim Wachstum. Es werden daher Baumarten gesucht, die eine hohe Toleranz gegen solche Umweltbedingungen haben. In einem Freilandversuch wurden auf einem Feld mit grossen lokalen Schwankungen des pH-Wert 120 Bäume einer bestimmten Art gepflanzt. Nach 3 Jahren wurde von jedem Baum die Höhe  $Y_i$  gemessen. Gleichzeitig war auch der pH-Wert  $x_i$  des Bodens an der entsprechenden Stelle bekannt. Die Daten können in einem Scatterplot dargestellt werden.



Das Hauptziel der Untersuchung besteht darin, den Zusammenhang zwischen der Ausgangsgrösse pH-Wert und der Zielvariable Baumhöhe anzugeben. Weiter soll angegeben werden, ob der pH-Wert die Baumhöhe überhaupt in signifikanter Art und Weise beeinflusst. Ebenfalls von Interesse ist die erwartete Baumhöhe, mit einer Art Vertrauensintervall, für einen bestimmten pH-Wert.

## 1.2 Modell

Der Zusammenhang zwischen einer erklärenden Variablen  $x$  und der Zielvariablen  $Y$  wird folgendermassen beschrieben:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ für alle } i = 1, \dots, n.$$

Dabei haben die einzelnen Grössen die folgende Bedeutung:

- $Y_i$  ist die *Zielvariable* der  $i$ -ten Beobachtung. Im vorliegenden Beispiel handelt es sich um die Höhe des  $i$ -ten Baums. Die Zielvariable ist eine Zufallsgrösse.
- $x_i$  ist die *erklärende Variable* der  $i$ -ten Beobachtung. Im vorliegenden Beispiel handelt es sich um den pH-Wert des Bodens am Standort des  $i$ -ten Baums. Die erklärende Variable wird als feste, nicht zufällige Grösse betrachtet.
- $\beta_0, \beta_1$  sind unbekannte Parameter, die sogenannten *Regressionskoeffizienten*. Diese sollen mit Hilfe der vorliegenden Beobachtungen geschätzt werden. Dabei ist  $\beta_0$  der so genannte *Achsenabschnitt* (engl. *Intercept*) und  $\beta_1$  die *Steigung* (engl. *Slope*). Letztere gibt an, um wie viel sich der Wert der Zielvariablen erhöht, wenn der  $x$ -Wert um eine Einheit zunimmt.
- $\varepsilon_i$  ist der *Restterm* oder *Fehler*. Es handelt sich um eine Zufallsgrösse, d.h. die Abweichung zwischen dem beobachteten Wert  $y_i$  und dem angepassten Wert auf der Gerade wird als zufällig interpretiert.

**Modellvoraussetzungen:** für den Fehler setzen wir voraus, dass der Erwartungswert  $E[\varepsilon_i] = 0$  ist, der Zusammenhang zwischen Ausgang- und Zielgrösse muss durch eine Gerade beschrieben werden. Weiter muss die Varianz  $Var(\varepsilon_i) = \sigma_\varepsilon^2$  konstant sein, und die Fehler dürfen keine Korrelation aufweisen, d.h.  $Cov(\varepsilon_i, \varepsilon_j) = 0$  für  $i \neq j$ .

Wir sprechen hier von einfacher linearer Regression, weil nur eine einzige erklärende Variable im Modell enthalten ist. Achtung: das Modell heisst nicht linear, weil eine Gerade angepasst wird, sondern weil die Modellgleichung linear in den Parametern  $\beta_0, \beta_1$  ist. So ist zum Beispiel auch  $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$  ein einfaches lineares Modell, während es sich bei  $Y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \varepsilon_i$  um ein nichtlineares Regressionsproblem handelt.

## 1.3 Schätzung

Hier stellen wir uns im Prinzip die Frage, welche Gerade am besten zu den  $n$  Punktpaaren  $(x_i, y_i)$  passt. Für jeden Punkt betrachten wir die vertikale Abweichung von der Geraden, das sogenannte *Residuum*

$$r_i = y_i - (\beta_0 + \beta_1 x_i)$$

Die Gerade soll nun so zu liegen kommen, dass die Summe der quadrierten Residuen so klein wie möglich ist. Man spricht auch von der *Methode der kleinsten Quadrate* (engl. *Least Squares*).

Das Minimierungsproblem kann entweder durch Nullsetzen der partiellen Ableitungen, oder auch durch geometrische Projektionsüberlegungen gelöst werden. Beachten sie die Analogie zur Ausgleichsrechnung in der Linearen Algebra. Wir stoßen auch hier wieder auf die Normalgleichungen

$$(X^T X)\beta = X^T y$$

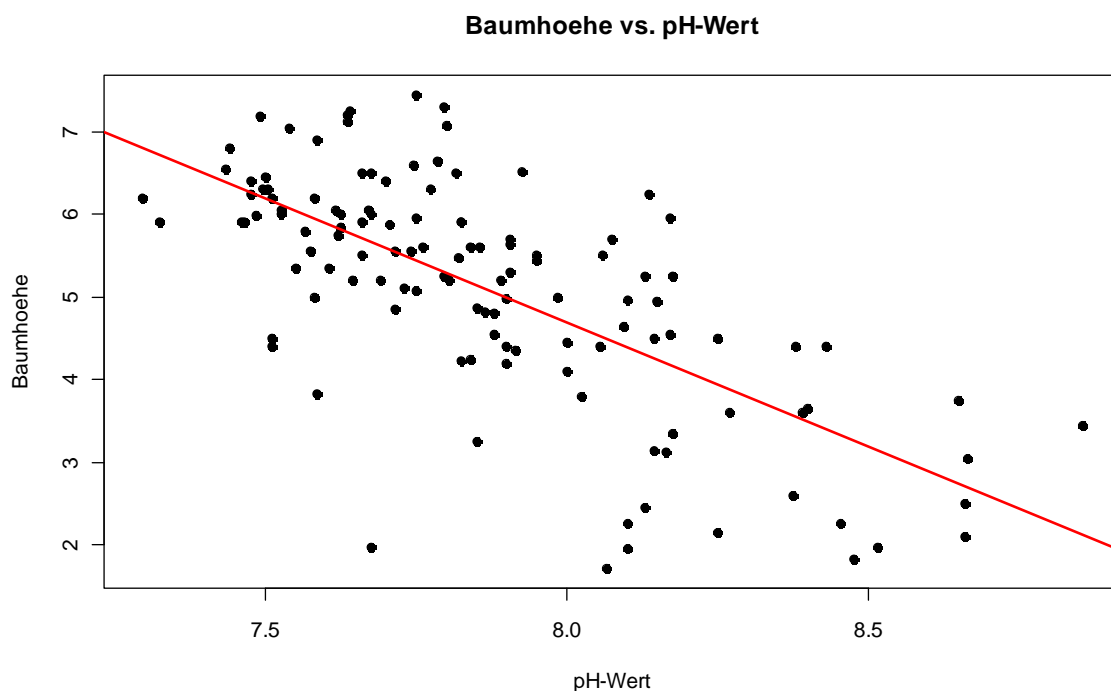
und erhalten als Lösung der Normalgleichungen die folgenden Schätzungen für die beiden Regressionsparameter:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{und} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Durch diese beiden Schätzungen ist die Regressionsgerade bestimmt. Sie ist gleich:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{für alle } i = 1, \dots, n.$$

und wir können sie im Scatterplot einzeichnen. Hierbei ist  $\hat{y}_i$  der vom Modell geschätzte Wert der Zielgröße. Man spricht auch vom *gefitteten Wert* (engl. *fitted* bzw. *predicted value*). Beachten sie, dass die Residuen der Differenz zwischen gefittetem und beobachtetem Wert entsprechen.



Natürlich ist es ein Stück weit willkürlich, gerade die Quadratsumme der Residuen zu minimieren. So könnte z.B. auch die Summe der absoluten Abweichungen, d.h. die  $L_1$ -Norm minimiert werden. Diese Norm, sowie einige weitere Minimierungskriterien, finden in der Praxis durchaus manchmal Anwendung. Sie zeichnen sich typischerweise durch grössere Robustheit gegenüber Ausreißern aus.

#### 1.4 Eigenschaften der Schätzungen

Gute Gründe sprechen hingegen auch für die Wahl der Kleinsten-Quadrate-Methode. Das *Gauss-Markov-Theorem* besagt, dass unter den Modellvoraussetzungen aus Kapitel 1.2 die Schätzungen  $\hat{\beta}_0, \hat{\beta}_1$  erwartungstreu sind (d.h.  $E[\hat{\beta}_0] = \beta_0$  und  $E[\hat{\beta}_1] = \beta_1$ ). Weiter haben sie unter allen erwartungstreuen, linearen Schätzern minimale Varianz, es sind also die genauesten Schätzungen. Man kann zeigen, dass

$$\text{Var}(\hat{\beta}_0) = \sigma_\varepsilon^2 \cdot \left( \frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right), \text{ und}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Dieses Resultat zeigt uns auch, wie wir durch ein geschicktes experimentelles Design genauere Schätzungen, d.h. eine „exakter bestimmte Regressionsgerade“ erhalten können.

- Wir können die Anzahl Beobachtungen  $n$  erhöhen.
- Wir können auf eine gute Streuung der  $x$ -Werte achten
- Wir können durch geeignete erklärende Variablen  $\sigma_\varepsilon^2$  klein halten
- Für  $\hat{\beta}_0$  hilft es, wenn der Mittelwert  $\bar{x}$  nahe bei null liegt.

#### 1.5 Schätzung von $\sigma_\varepsilon^2$

Neben den Regressionskoeffizienten ist auch noch die *Varianz der zufälligen Fehler* zu schätzen, die wir für alle möglichen Tests und Vertrauensintervalle benötigen. Sie basiert auf der *Residuenquadratsumme* ( $RSS$ , engl. *Residual Sum of Squares*) und lautet:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## 1.6 Tests und Vertrauensintervalle

Bis zu diesem Punkt haben wir nur Annahmen über Erwartungswert, Varianz und Korrelation der zufälligen Fehler gemacht, jedoch keine bestimmte Verteilung vorausgesetzt. Das bedeutet konkret, dass die obigen Resultate unabhängig von der Verteilung gelten.

In diesem Abschnitt wollen wir nun testen, ob die erklärende Variable  $x$  einen signifikanten Einfluss auf die Zielgröße  $Y$  hat. Dies ist nur möglich, wenn wir eine Verteilungsannahme für die zufälligen Fehler treffen. Wir setzen also im folgenden voraus, dass

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2), \text{ i.i.d..}$$

Beachten sie, dass diese Annahme mit den Methoden aus Kapitel 1.9 überprüft werden muss, wenn man sich auf die im Folgenden vorgestellten Tests verlassen will. Sind die Voraussetzungen nicht erfüllt, so kann es zu groben Fehlschlüssen kommen.

Um zu entscheiden, ob die erklärende Variable  $x$  einen signifikanten Einfluss auf die Zielgröße  $Y$  hat, testet man die Nullhypothese  $H_0: \beta_1 = 0$  gegen die Alternative  $H_A: \beta_1 \neq 0$ . Als Testgröße verwenden wir

$$T = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sqrt{\text{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}_\varepsilon^2 / \sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Die Testgröße hat eine  $t$ -Verteilung mit  $n-2$  Freiheitsgraden. Diese wird zur Bestimmung von Verwerfungs- und Annahmehereich, bzw. für die Bestimmung des  $p$ -Werts verwendet. Verwirft man die Nullhypothese, so wird der Zusammenhang zwischen erklärender Variable und Zielgröße als statistisch gesichert betrachtet, d.h. die Steigung ist signifikant von null verschieden. Will man prüfen, ob sich der Achsenabschnitt  $\hat{\beta}_0$  signifikant von null unterscheidet, so geht man auf analoge Weise vor.

## 1.7 Output von Statistikpaketen

Wird die Regressionsanalyse mit einem Statistikpaket ausgeführt, so werden im Output nicht nur die Punktschätzungen für  $\beta_0, \beta_1$  angegeben (Spalte „*Estimate*“), sondern in der Regel auch deren Standardabweichungen (Spalte „*Std. Error*“), der Wert der Testgröße  $T$  (Spalte „*t value*“), sowie der  $p$ -Wert zu den jeweiligen Nullhypothesen (Spalte „*Pr(>|t|)*“).

```
> summary(fit)
Call:
lm(formula = height ~ ph, data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-3.70195 -0.54712  0.08745  0.66626  2.00330
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	28.7227	2.2395	12.82	<2e-16	***
ph	-3.0034	0.2844	-10.56	<2e-16	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.008 on 121 degrees of freedom

Multiple R-squared: 0.4797, Adjusted R-squared: 0.4754

F-statistic: 111.5 on 1 and 121 DF, p-value: < 2.2e-16

Weiter ist auch die Punktschätzung für  $\sigma_\varepsilon^2$  angegeben („*Residual standard error*“), mit der zugehörigen Anzahl Freiheitsgrade  $n-2$  („*degrees of freedom*“), aus welcher sich direkt die Anzahl Beobachtungen ablesen lässt, mit welcher die Regression gerechnet wurde. Zusätzlich wird auch noch die Grösse *Multiple R-squared* angegeben. Sie berechnet sich als

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \in [0,1]$$

und gibt den Anteil der durch die Regression erklärten Varianz an der Gesamtvarianz an.

## 1.8 Prognose

Die geschätzten Parameter, bzw. die angepasste Regressionsgerade kann nun benutzt werden, um für ein beliebiges  $x^*$  den Wert der Zielvariablen vorherzusagen. Wir verwenden dazu einfach:

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

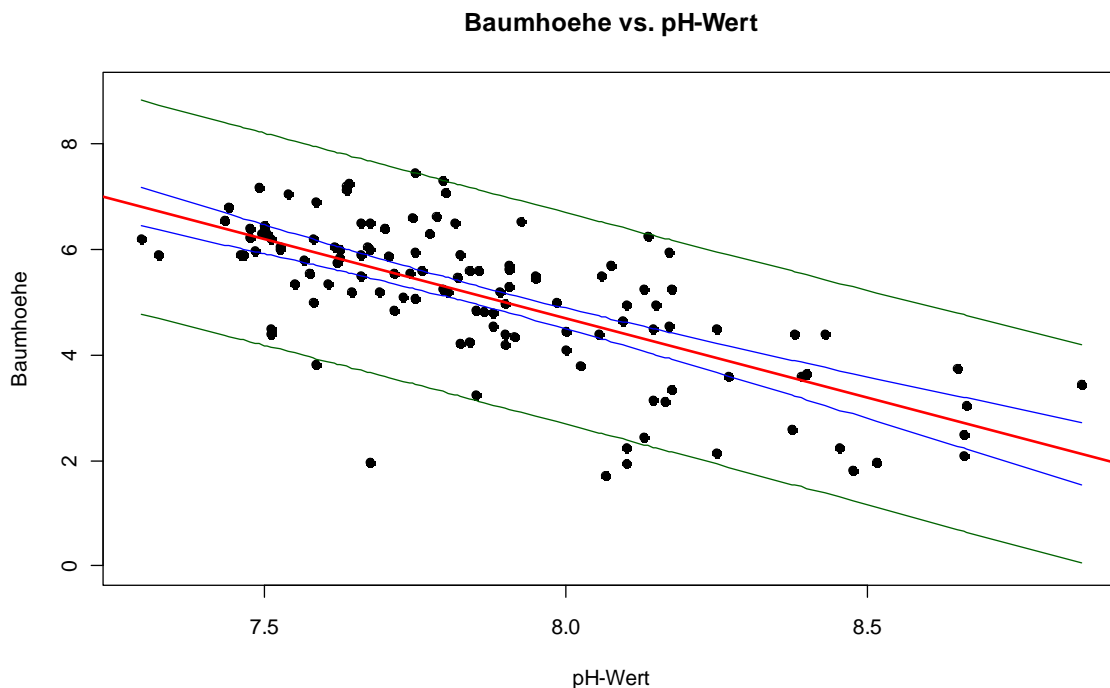
Hierbei ist zu berücksichtigen, dass normalerweise nur Vorhersagen innerhalb des Bereichs der für die Anpassung der Gerade verwendeten  $x$ -Werte verlässlich sind. Man spricht in diesem Fall von *Interpolation*. Hingegen sind *Extrapolationen*, über den Bereich der verwendeten  $x$ -Werte hinaus, generell mit Vorsicht zu geniessen.

**Beispiel:** Für eine pH-Wert von 8.0 erwarten wir eine Baumhöhe von  $28.7227 + (-3.0034 \cdot 8.0) = 4.4955$  Metern. Hingegen ist es nicht sinnvoll, mit Hilfe der Regressionsgerade die Baumhöhe für einen pH-Wert von 5.0 anzugeben.

Wir können nun ein 95%-Vertrauensintervall für den Vorhersagewert  $\hat{y}^*$  angeben. Es ist folgendermassen bestimmt:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{0.975;n-2} \cdot \sigma_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Wir können dieses Intervall für beliebige  $x^*$  angeben, und in der folgenden Grafik in blauer Farbe als *Vertrauensbereich* für die angepasste Gerade einzeichnen. Dieser ist, wie aufgrund der Gleichung leicht sichtbar, in der Mitte schmäler als an den Rändern.



Mit dem Vertrauensbereich für die Regressionsgerade, das ja aus Vertrauensintervallen für die angepassten Werte bestimmt war, ist jedoch noch nicht klar, wo eine neue Beobachtung zu liegen kommt. Die einzelnen Beobachtungen streuen ja noch zusätzlich um den erwarteten Wert herum. Das Prognoseintervall für  $y^*$  ist gegeben durch:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{0,975;n-2} \cdot \sigma_\varepsilon \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Wiederum können wir das Prognoseintervall für alle  $x^*$  bestimmen und als *Prognosebereich* mit dunkelgrüner Farbe im Scatterplot einzeichnen.

### 1.9 Prüfen der Voraussetzungen: Residuenanalyse

Nach der Anpassung einer Regressionsgeraden soll überprüft werden, ob die Modellvoraussetzungen aus Kapitel 1.2 und die Normalverteilungsannahme erfüllt, und die Resultate somit gültig sind. Zu prüfen ist:

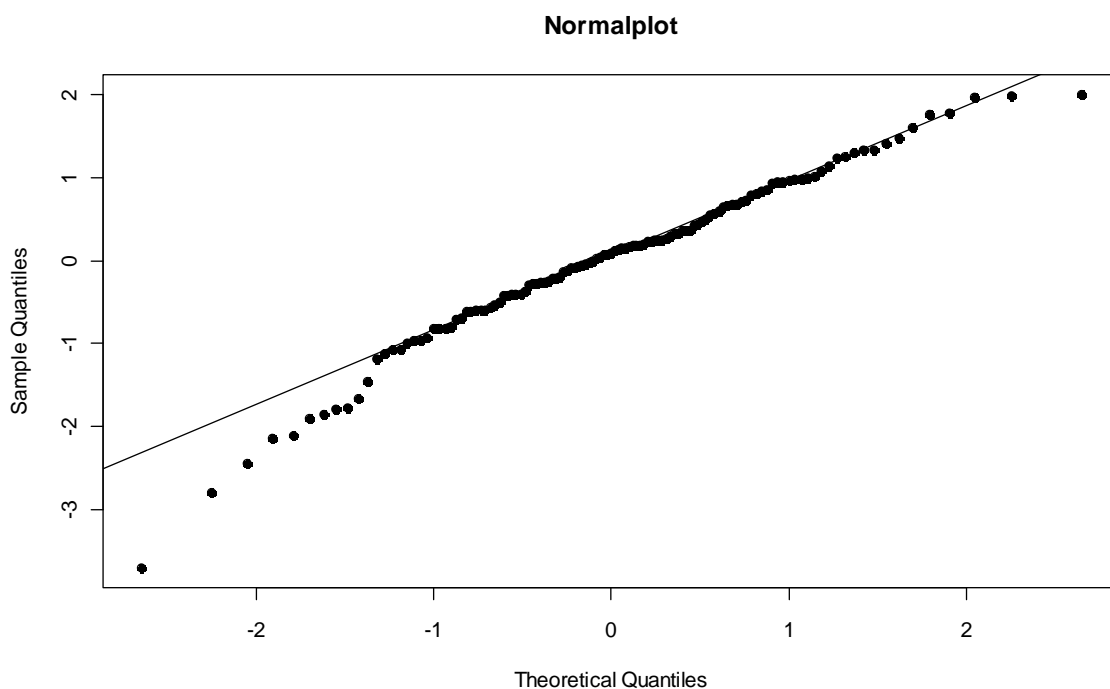
- Der Zusammenhang zwischen  $x$  und  $Y$  ist genähert linear, d.h. der Erwartungswert der Fehler  $\varepsilon_i$  ist auf dem ganzen Wertebereich null.
- Die Fehler  $\varepsilon_i$  haben eine konstante Streuung  $\sigma_\varepsilon^2$  und sind zudem (zeitlich) unkorreliert.

- Falls Tests oder Vertrauens-/Prognosebereiche berechnet werden, so müssen die Fehler  $\varepsilon_i$  zusätzlich auch normalverteilt sein.

Die Überprüfung findet mit grafischen Methoden statt. Dabei werden die Residuen  $r_i$  gegen verschiedene andere Variablen dargestellt. Hier werden die beiden wichtigsten Plots vorgestellt.

### Normalplot

Die Annahme der Normalverteilung wird mit dem sogenannten *Normalplot* überprüft. Dabei werden die nach Grösse geordneten Residuen gegen die entsprechenden Quantile der Normalverteilung geplottet. Wenn die Fehler  $\varepsilon_i$  normalverteilt sind, so gilt dies auch für die Residuen. Die Punkte im Normalplot sollten also entlang einer Gerade liegen.



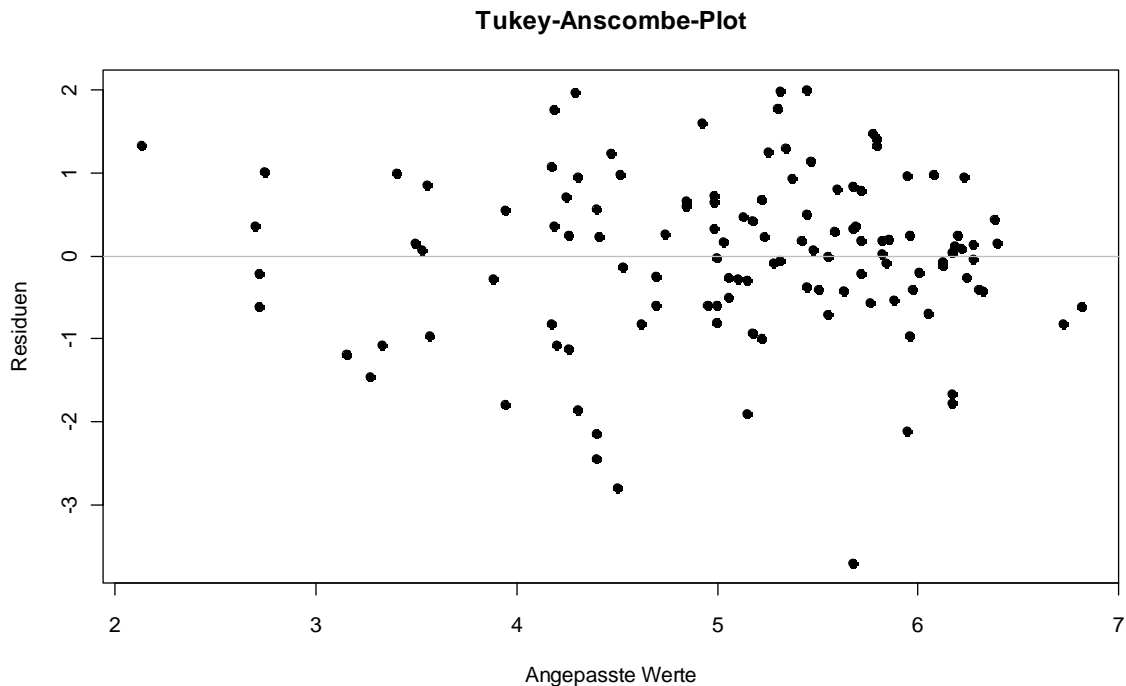
**Beispiel:** Der Normalplot für das Beispiel der Baumhöhen zeigt, dass vor allem gegen unten einige Abweichungen vorhanden sind, die etwas grösser sind, als dies eine Normalverteilung der Fehler suggerieren würde. Jedoch kann die in diesem Beispiel beobachtete Abweichung gerade noch als tolerierbar erachtet werden.

Würde der Normalplot eine rechtsschiefe Verteilung zeigen, so kann eine Logarithmus- oder Wurzeltransformation der Zielgrösse Abhilfe schaffen. Zeigt der Normalplot einzelne Beobachtungen mit betragsmässig grossen Residuen, so soll abgeklärt werden, ob es sich hierbei um grobe (Abschreib-)fehler etc. handelt. Falls ja, so können diese Beobachtungen entfernt oder korrigiert, und die Analyse wiederholt werden.

Für andere, systematische Abweichungen von der Normalverteilungsannahme sind Methoden nötig, die über den Rahmen einer Einführungsvorlesung hinausgehen. Konsultieren sie falls nötig einen Spezialisten!

## Tukey-Anscombe-Plot

Mit dieser Darstellung können sowohl Abweichungen von der Linearität, als auch nichtkonstante Varianz entdeckt werden. Man trägt auf der x-Achse die angepassten Werte, und auf der y-Achse die Residuen auf. Im Idealfall befinden sich alle Residuen in einem horizontalen Band mit konstanter Breite, und streuen zufällig.



**Beispiel:** Der Tukey-Anscombe-Plot für das Baumhöhen-Beispiel zeigt keine groben Verletzungen der Modellvoraussetzungen auf. Es fällt einzig, wie schon im Normalplot, ein Ausreisser gegen unten auf. Es wäre sicher zu prüfen, ob sich hier kein Abschreibfehler oder dergleichen eingeschlichen hat.

Bei nichtkonstanter Streuung kann eine Transformation Abhilfe schaffen, oder es muss gewichtete Regression (die hier nicht besprochen wird) verwendet werden. Auch bei Nichtlinearität können Transformationen hilfreich sein, oder es müssen zusätzliche Variablen ins Modell aufgenommen werden.

### 1.10 Ausblick: Multiple Regression

Das Modell der einfachen linearen Regression lässt sich auf den Fall verallgemeinern, wo die Zielvariable  $Y$  nicht nur von einer, sondern von mehreren erklärenden Variablen  $x^{(1)}, x^{(2)}, \dots, x^{(p)}$  beeinflusst wird. Die Regressionsgleichung lautet dann:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \dots + \beta_p x_i^{(p)} + \varepsilon_i \text{ für alle } i = 1, \dots, n.$$

Über die zufälligen Fehler werden hier wieder genau dieselben Annahmen wie im Kapitel 1.2 getroffen. Wichtig ist zu berücksichtigen, dass das Modell der multiplen Regression mehr Information liefert als das Ausführen einer Sequenz von

einfachen Regressionen mit allen enthaltenen Variablen. Die multiple Regression kann nämlich das Zusammenspielen der verschiedenen Einflussfaktoren berücksichtigen.

**Beispiel:** Bei den Baumhöhen war nicht nur der pH-Wert des Bodens verfügbar, sondern auch die „Sodium Absorption Ratio“, welche einen etwas anderen Aspekt der Basizität erfasst. Erklärt man die Baumhöhe mit beiden Variablen gleichzeitig, so ist eine etwas genauere Beschreibung möglich.

Die geschätzten Parameter  $\hat{\beta}_1, \dots, \hat{\beta}_p$  sowie  $\hat{\sigma}_\varepsilon^2$  erhält man wieder durch Anwendung der Methode der Kleinsten Quadrate, welche auch hier zu den Normalgleichungen führen. Bei der Verwendung von Software-Paketen zur Regressionsrechnung ist die Erweiterung zu multipler Regression meist „straightforward“. Im Output werden weiterhin Schätzwerte, Standardfehler, Testgrößen und p-Werte für die Nullhypothesen  $H_0: \beta_j = 0$  für alle  $j = 0, \dots, p$  geliefert.

Die Berechnung von Vertrauens-/und Prognosebereich, sowie die Residuenanalyse spielt sich sehr ähnlich wie im Fall der einfachen linearen Regression ab.

### 1.11 Ausblick: Verallgemeinerte lineare Modelle

Für die bisher im Kapitel 1 besprochenen Methoden gingen wir von der Annahme aus, dass die Zielvariable  $Y$  eine stetige Zufallsgröße ist. Dies ist nicht bei allen Regressionsproblemen der Fall. Es kann z.B. auch interessant sein, das Überleben eines Versuchstiers in Abhängigkeit von erklärenden Größen zu untersuchen.

Hierzu passt weder das Modell der einfachen, noch der multiplen linearen Regression, sondern es ist eine logistische Regression gefragt, die ihrerseits den verallgemeinerten linearen Modellen zugeordnet ist. Falls sie vor einem solchen Problem stehen, bilden sie sich weiter, oder konsultieren sie einen Spezialisten.