Applied Statistical Regression
HS 2011 – Week 01

Marcel Dettling

Institute für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, September 26, 2011

Applied Statistical Regression
HS 2011 – Week 01

Your Lecturer

Name:	Marcel Dettling
Age:	36 Jahre
Civil Status:	Married, 2 children
Education:	Dr. Math. ETH
Position:	Lecturer at ETH Zürich and ZHAW Project Manager R&D at IDP, a ZHAW institute
Hobbies:	Rock climbing, Skitouring, Paragliding, …

Applied Statistical Regression
HS 2011 – Week 01

Course Organization

Applied Statistical Regression – HS 2011

People:

Lecturer: **Dr. Marcel Dettling** (marcel.dettling@zhaw.ch)

Coordinators: **Christian Kerkhoff** (kerkhoff@stat.math.ethz.ch)
Philipp Rütimann (rutimann@stat.math.ethz.ch)

Course Schedule:

All lectures will be held at HG D1.1, on Mondays from 8.15-9.00, resp. 9.15-10.00.

Week	Date	L/E	Topics
01	19.09.2011	---	---
02	26.09.2011	L/L	Simple regression
03	03.10.2011	E/E	Introduction to R
04	10.10.2011	L/L	Multiple regression
05	17.10.2011	L/E	Model diagnostics
06	24.10.2011	L/L	Model extensions
07	31.10.2011	L/E	Model choice 1
08	07.11.2011	L/L	Model choice 2
09	14.11.2011	L/E	Introduction to GLMs
10	21.11.2011	L/L	Logistic regression
11	28.11.2011	L/E	Regression for count data
12	05.12.2011	L/L	Regression for nominal and ordinal response
13	12.12.2011	E/E	Exercises
14	19.12.2011	L/L	Advanced Topics

Exercise Schedule:

The exercises start on October 3, 2011 from 8.15 to 10.00 with an introduction to the statistical software package R. The location of this R-introduction is to be announced. Thereafter, the exercise schedule is as follows:

Series	Date	Topic	Hand-In	Discussion
01	03.10.2011	Data analysis with R	---	03.10.2011
02	03.10.2011	Simple linear regression	10.10.2011	17.10.2011
03	17.10.2011	Multiple regression/diagnostics	24.10.2011	31.10.2011
04	31.10.2011	Multiple regression/various	07.11.2011	14.11.2011
05	14.11.2011	Model choice	21.11.2011	28.11.2011
06	28.11.2011	Logistic regression	05.12.2011	12.12.2011
07	12.12.2011	Count and ordinal data	---	12.12.2011

All exercises except the first one take place at HG E41 (group of Kerkhoff) and HG D1.1 (group of Rütimann). All students whose last name starts with letters A-K visit the group of Kerkhoff, whereas the ones with letters L-Z visit the Rütimann group.

The solved exercises should be placed in the corresponding tray in HG J68 until 11.55am of the due date. They can also be sent via e-mail to the respective assistant. Please note that only recapitulatory documents shall be handed in, but no R script files.

Introduction to Regression

Everyday question:

How does a target (value) of special interest depend on several other (explanatory) factors or causes.

Examples:

- growth of plants, affected by fertilizer, soil quality, …
- apartment rents, affected by size, location, furnishment, …
- airplane fuel consumption, affected by tow, distance, weather, …

Regression:

- quantitatively describes relation between predictors and target
- high importance, most widely used statistical methodology

The Linear Model

Simple and appealing way for describing predictor/target relation!

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

For specifying this model, we need to estimate its parameters. In order to do so, we need data. Usually, we are given *n* data points.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

Estimation is such that the errors are “small”, i.e. such that the sum of squared residuals is minimized. Some additional assumption are necessary, too.

Goals with Linear Modeling

Goal 1: ***To understand the causal relation, doing inference***

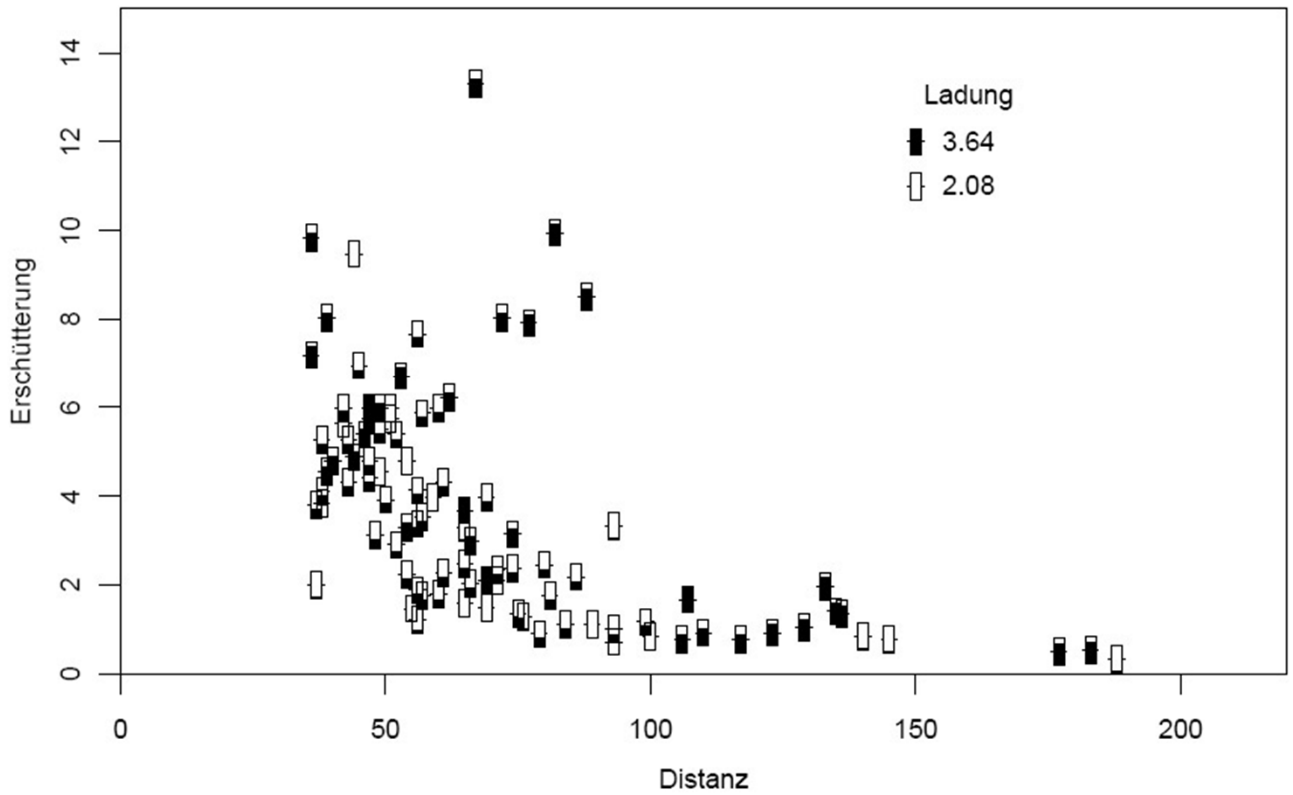
- Does the fertilizer positively affect plant growth?
- Regression is a tool to give an answer on this
- However, showing causality is a different matter

Goal 2: ***Target value prediction for new explanatory variables***

- How much fuel is needed for the next flight?
- Regression analysis formalizes “prior experience”
- It also provides an idea on the uncertainty of the prediction

Versatility of Linear Modeling

"Only" linear models: is that a problem? → **NO**



Applied Statistical Regression
HS 2011 – Week 01

Topics of the Course

- 01 - Introduction
- 02 - Simple Linear Regression
- 03 - Multiple Linear Regression
- 04 - Extending the Linear Model
- 05 - Model Choice
- 06 - Generalized Linear Models
- 07 - Logistic Regression
- 08 - Nominal and Ordinal Response
- 09 - Regression with Count Data
- 10 - Modern Regression Techniques

Synopsis: What will you learn?

Over the entire course, we try to address the questions:

- *Is a regression analysis the right way to go with my data?*
- *How to estimate parameters and their confidence intervals?*
- *What assumptions are behind, and when are they met?*
- *Does my model fit? What can I improve it it does not?*
- *How can identify the “best” model, and how to choose it?*

Before You Start…

> The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill.
>
> *Albert Einstein*

Process:

1) Understand and formulate the problem
2) Obtain the data and check them
3) Do a technically correct analysis
4) Draw conclusions

it's an iterative process!

Common Mistakes

The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill.

Albert Einstein

Though it shall be avoided at any cost, it happens again:

- Thoughtless collecting of data, without a clear question
- Statistical analyses without having a precise goal/question
- One just reports what was found by coincidence

→ *Act better!*

Good Practice in Data Analysis

1) Try to understand the background. Take the time to acquire knowledge on the subject.

2) Make sure that the question is precisely formulated. This often requires some awkward begging on your partners, because they don't know exactly themselves. But it's worth it!

3) Avoid "fishing expeditions", where you search your data until you have found "something". Finally, there is always something standing out. However, it's often just random variation or artefacts.

Good Practice in Data Analysis

4) Choose an appropriate amount of complexity. Sophisticated methodology should not be used for vanity reasons, but only if it is really required.

5) Try to translate the question from the applied field into the world of statistics, i.e. clearly indicate, which statistical analyses answer what question(s) how precisely.

   - that's not simple!
   - it cannot be done automatically!
   - education and having the knowledge is key!

Garbage In, Garbage Out

IMPORTANT:

Feeding some data into some statistical method, make it run without obtaining and error message and producing some output is one thing…

Withouth a thoughtful approach, such results are usually worthless for yourself and your partners. Thus, be critical: both against yourself, as well as against third party analyses.

The Data

Origin of the data:

- Are you working with experimental or observational data?
  Is it a thought-about sample, or is it a convenience sample?
  In both latter cases, be careful!

→ The origin of the data has a strong impact on the quality of your findings, and on the conclusions that can be drawn.

→ If the sample is not representative: all warnings regarding the results are quickly forgotten, and one tends to only remember what is nice and shiny!

The Data

Non-Response – systematically missing values

- *Is there non-response, i.e. systematically missing values?* Are there some particular configurations where the measurements "couldn't be made", or are there typical groups of people who did not respond, etc.?

→ These missing data are often equally important as the ones which are present, i.e. they also have a message.

→ In such cases, goals and conclusions often need to be revised, as there are cases/things we could not observe.

The Data

Coding of the variables

- Be careful on how non-response and randomly missing data are coded! Always and only use "NA" for this.
- Are categorical variables correctly represented, and cannot be falsely interpreted as numeric values?
- For numerical varlues: are the measurement units correct and sensible, such that an analysis or comparison is possible?
- In real data, at least if they have a certain size, there are almost always some gross errors. Be careful in this respect, and make corrections where necessary.

Simple Linear Regression

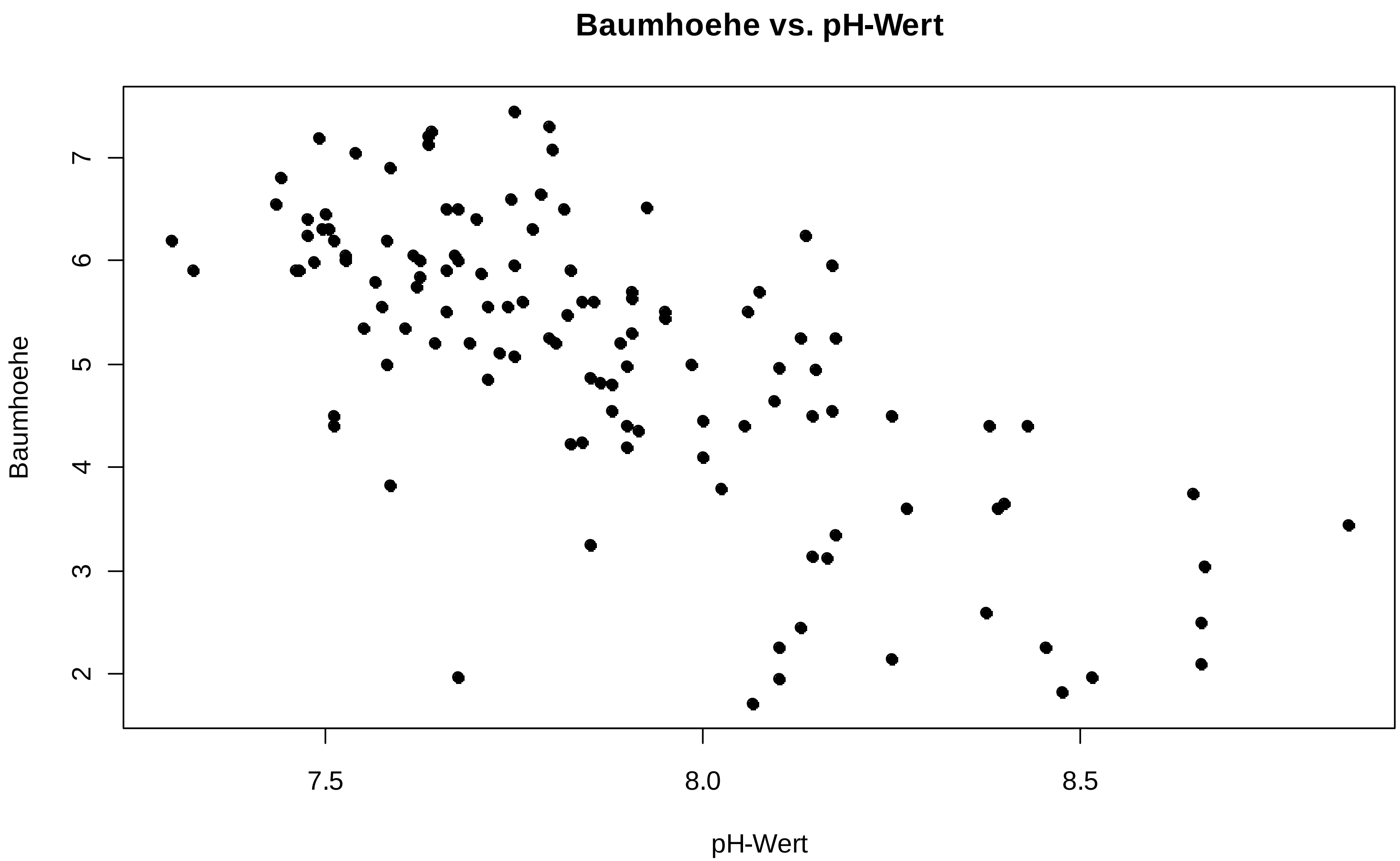
Example:

In India, it was observed that alkaline soil hampers plant growth. This gave rise to a search for tree species which show high tolerance against these conditions.

An outdoor trial was performed, where 120 trees of a particular species were planted on a big field with considerable soil pH-value variation.

After 3 years of growth, every trees height was measured. Additionally, the pH-value of the soil in the vicinity of each tree was determined and recorded.

Scatterplot: Tree Height vs. pH-value



The Simple Linear Regression Model

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for all i=1,…,n

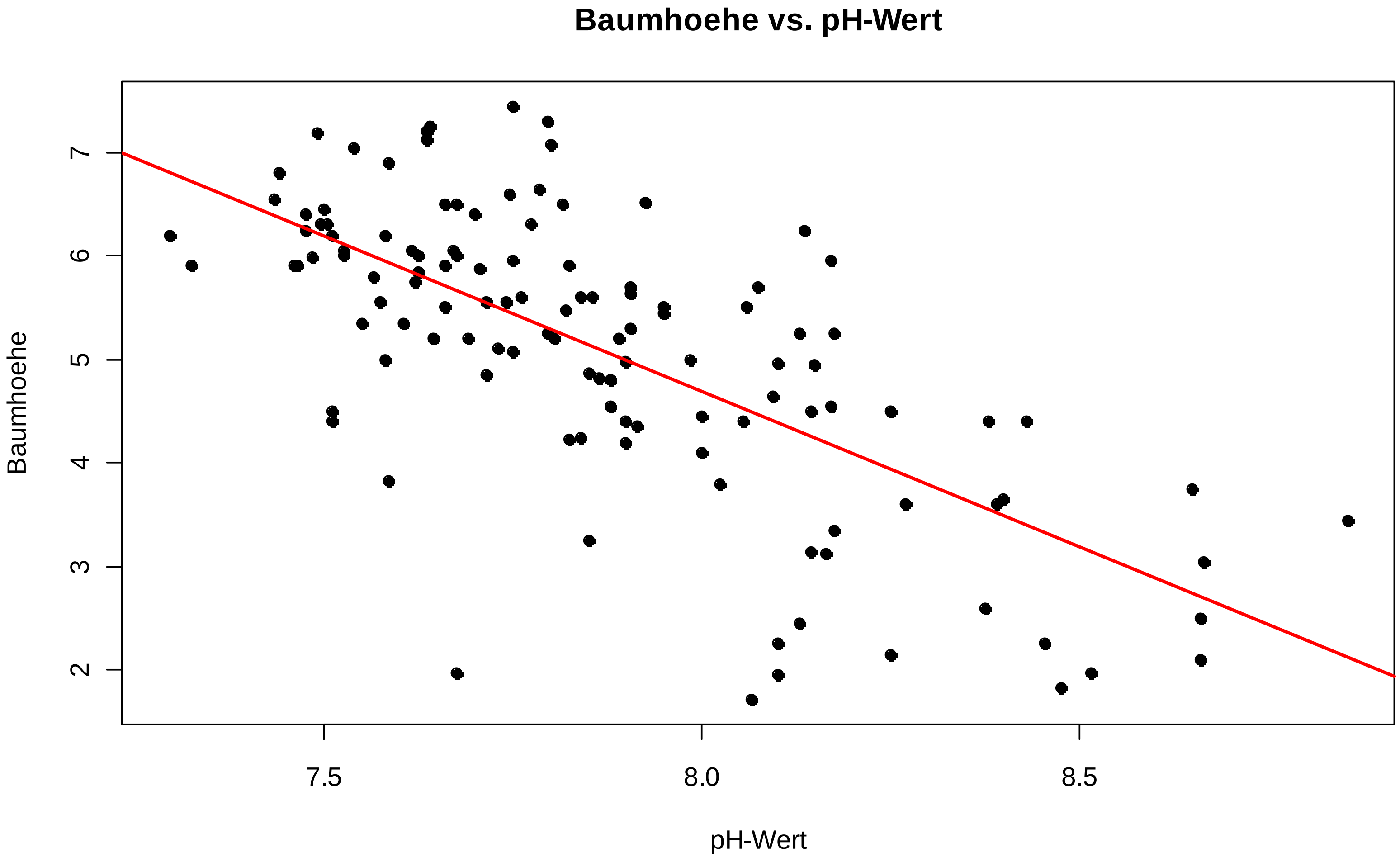
→ What is the meaning of the parameters?
- response/predictors
- regression coefficients
- error term

→ Which assumptions are made (for the error term)?
- zero expectation
- constant variance
- uncorrelated
- but nothing (yet) on the distribution!

Parameter Estimation

→ **See blackboard…**

Regression Line



Gauss-Markov-Theorem

And: what can be done to obtain better estimates?

→ **See blackboard…**

Estimation of the Error Variance

Besides the regression coefficients, we also need to estimate the error variance. We require it for doing inference on the estimated parameters. The estimate is based on the *residual sum of squares* (abbreviation: RSS), in particular:

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

This is (almost) the “usual” variance estimator!

Inference on the Parameters

Goal: is the relation target/predictor statistically significant?

→ For this, we need: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, i.i.d.

The test setup has the following hypotheses:

→ $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

Test statistic:

→ $$T = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sqrt{Var(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}_\varepsilon^2 \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2}} \sim t_{n-2}$$

Output of Statistical Software Packages

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

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 ***

Residual stand. err.: 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
```

Prediction

The regression line can now be used for predicting the target value at an arbitrary (new) value. We simply do as follows:

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

Example: For a pH-value of 8.0, we expect a tree height of

$$28.7227 + (-3.0034 \cdot 8.0) = 4.4955$$

A word of caution:

Doing interpolation is usually fine, but extrapolation (i.e. giving the tree height for pH-value 5.0) is generally “dangerous”.

Confidence and Prediction Intervals

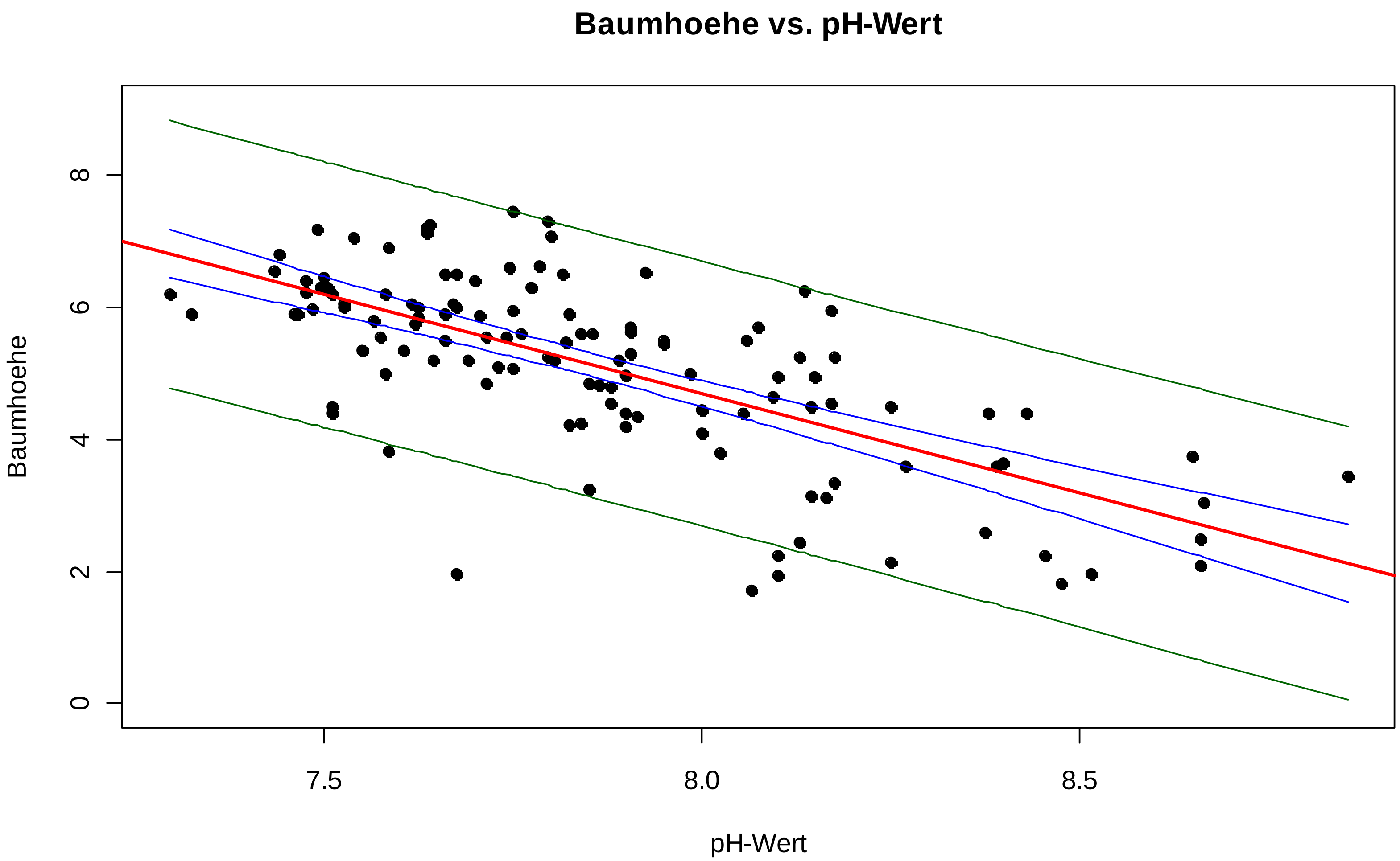
95% confidence interval: this is for the fitted value!

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

95% prediction interval: this is for future observations!

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

Confidence and Prediction Intervals



Does the Regression Line Fit Well?

If not, we are bound to incorrect conclusions!!!

Thus, it's wise to check the following:

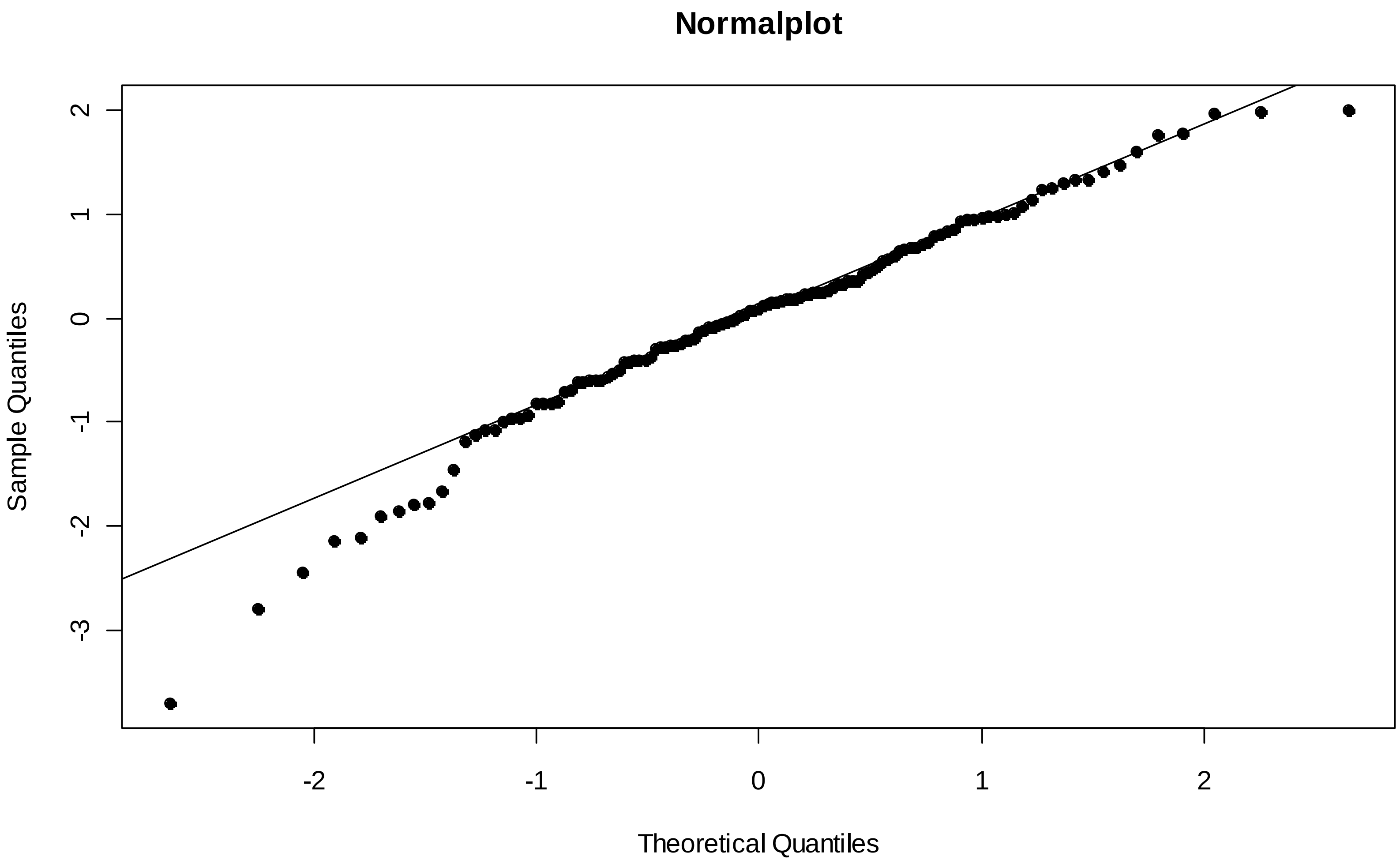
- regression line is the correct relation, zero error expected
  → Residuals vs. predictor, or Tukey-Anscombe plot

- scatter is constant, and the residuals are uncorrelated
  → Residuals vs. predictor, or Tukey-Anscombe plot

- errors/residuals are normally distributed
  → Normal plot of the residuals

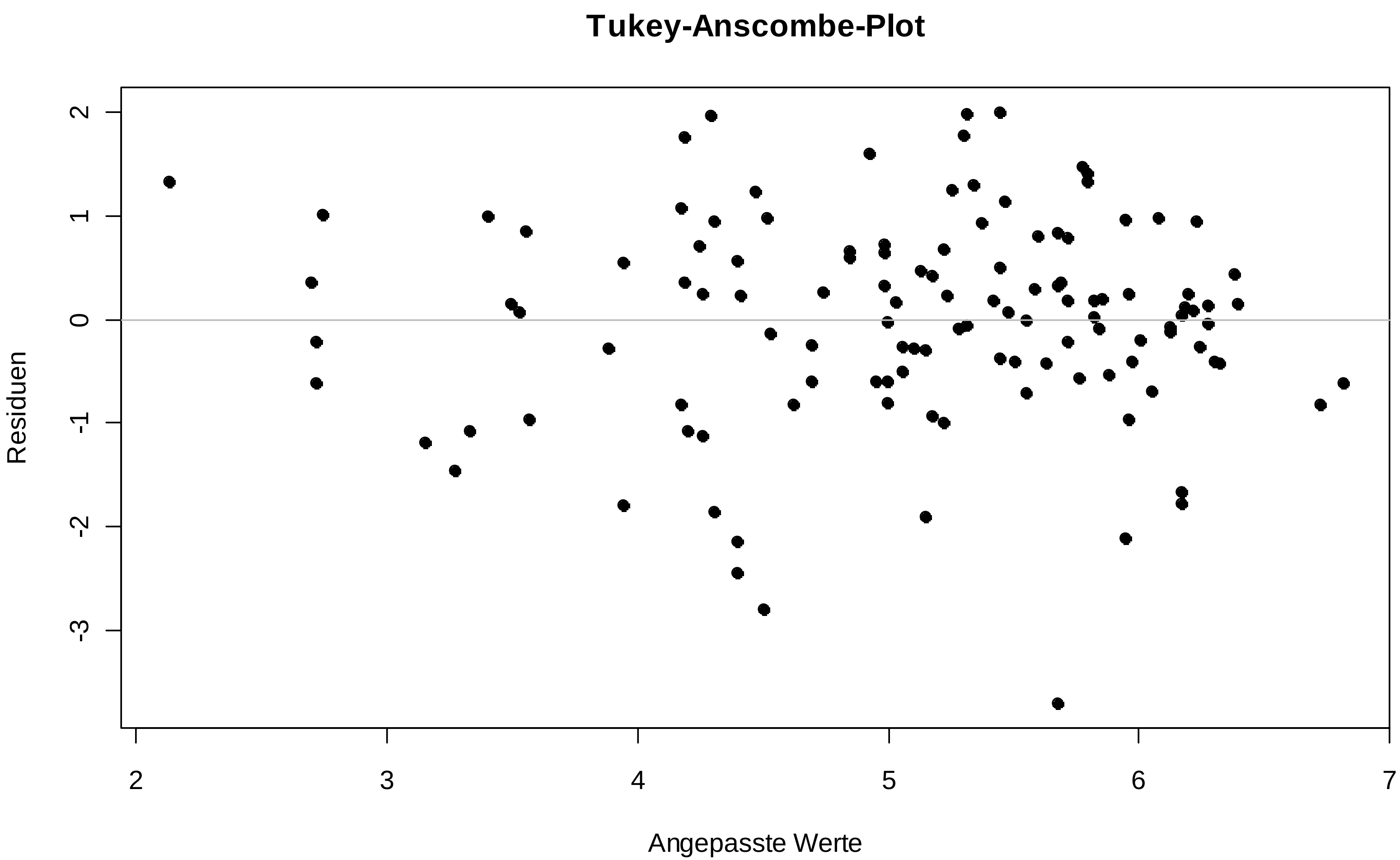
Applied Statistical Regression
HS 2011 – Week 01

Normal Plot



Applied Statistical Regression
HS 2011 – Week 01

Tukey-Anscombe Plot



How to Deal with Violations?

- A few gross outliers
  → check them for errors, correct or omit

- Prominent long-tailed distribution
  → robust fitting, to be discussed later

- Skewed distribution and/or non-constant variance
  → log- or square-root-transform the response
  → use a different model (generalized linear model)

- Non-random structure in the Tukey-Anscombe plot
  → improve the model, i.e. predictors are missing