Applied Statistical Regression
HS 2010 – 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 27, 2010

Applied Statistical Regression
HS 2010 – Week 01

Your Lecturer

Name: Marcel Dettling

Education: Dr. Math. ETH

Job: Project Manager R&D @ ZHAW
Lecturer @ ETH Zürich & ZHAW

Private:

Applied Statistical Regression
HS 2010 – Week 01

Course Organization

Applied Statistical Regression – HS 2010

People:

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

Coordinators: **Christian Kerkhoff** (kerkhoff@stat.math.ethz.ch)
Fabio Sigrist (sigrist@stat.math.ethz.ch)

Course Schedule:

All lectures will be held at HG D3.2, on Mondays from 8.15-9.00, resp. 9.15-10.00.

Week	Date	L/E	Topics
01	20.09.2010	---	---
02	27.09.2010	L/L	Simple regression
03	04.10.2010	E/E	Introduction to R
04	11.10.2010	L/L	Multiple regression
05	18.10.2010	L/E	Model diagnostics
06	25.10.2010	L/L	Model extensions
07	01.11.2010	L/E	Model choice 1
08	08.11.2010	L/L	Model choice 2
09	15.11.2010	L/E	Introduction to GLMs
10	22.11.2010	L/L	Logistic regression
11	29.11.2010	L/E	Regression for count data
12	06.12.2010	L/L	Regression for nominal and ordinal response
13	13.12.2010	E/E	Exercises
14	20.12.2010	L/L	Advanced Topics

Exercise Schedule:

The exercises start on October 4, 2010 from 8.15 to 10.00 with an introduction to the statistical software package R. Location of this R-introduction: to be announced. Thereafter, the exercise schedule is as follows:

Series	Date	Topic	Hand-In	Discussion
01	04.10.2010	Data analysis with R	---	04.10.2010
02	04.10.2010	Simple linear regression	11.10.2010	18.10.2010
03	18.10.2010	Multiple regression/diagnostics	25.10.2010	01.11.2010
04	01.11.2010	Multiple regression/various	08.11.2010	15.11.2010
05	15.11.2010	Model choice	22.11.2010	29.11.2010
06	29.11.2010	Logistic regression	06.12.2010	13.12.2010
07	13.12.2010	Count and ordinal data	---	13.12.2010

All exercises except the first one take place at HG D3.2 (group of Kerkhoff) and HG D1.2 (group of Sigrist). All students whose last name starts with letters A-K visit the group of Kerkhoff, whereas the ones with letters L-Z visit the Sigrist 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.

Applied Statistical Regression
HS 2010 – Week 01

Introduction

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!

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

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.

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

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

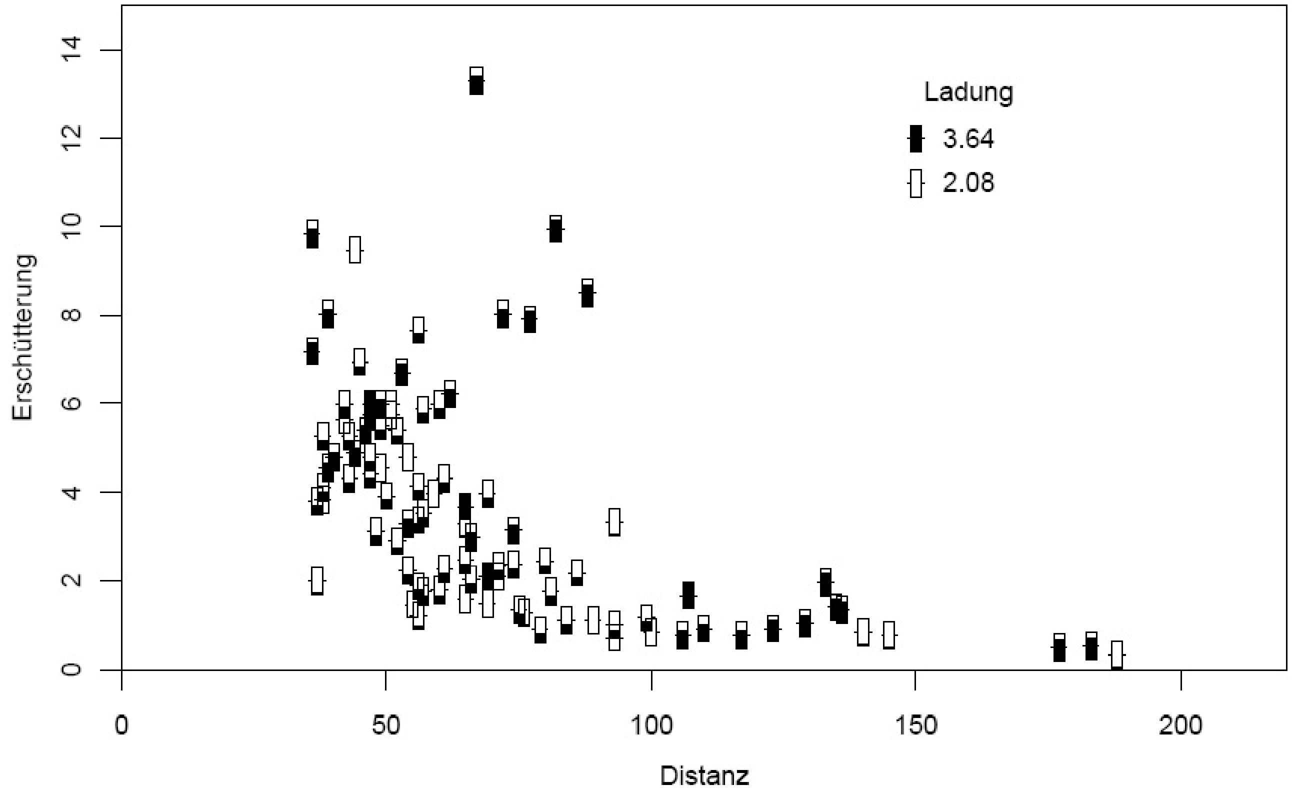
Goal 2: ***Target value prediction for new explanatory variables***

- 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

Applied Statistical Regression
HS 2010 – Week 01

Versatility of Linear Modeling

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



Applied Statistical Regression
HS 2010 – 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

Applied Statistical Regression
HS 2010 – Week 01

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?*

Simple Linear Regression

Example:

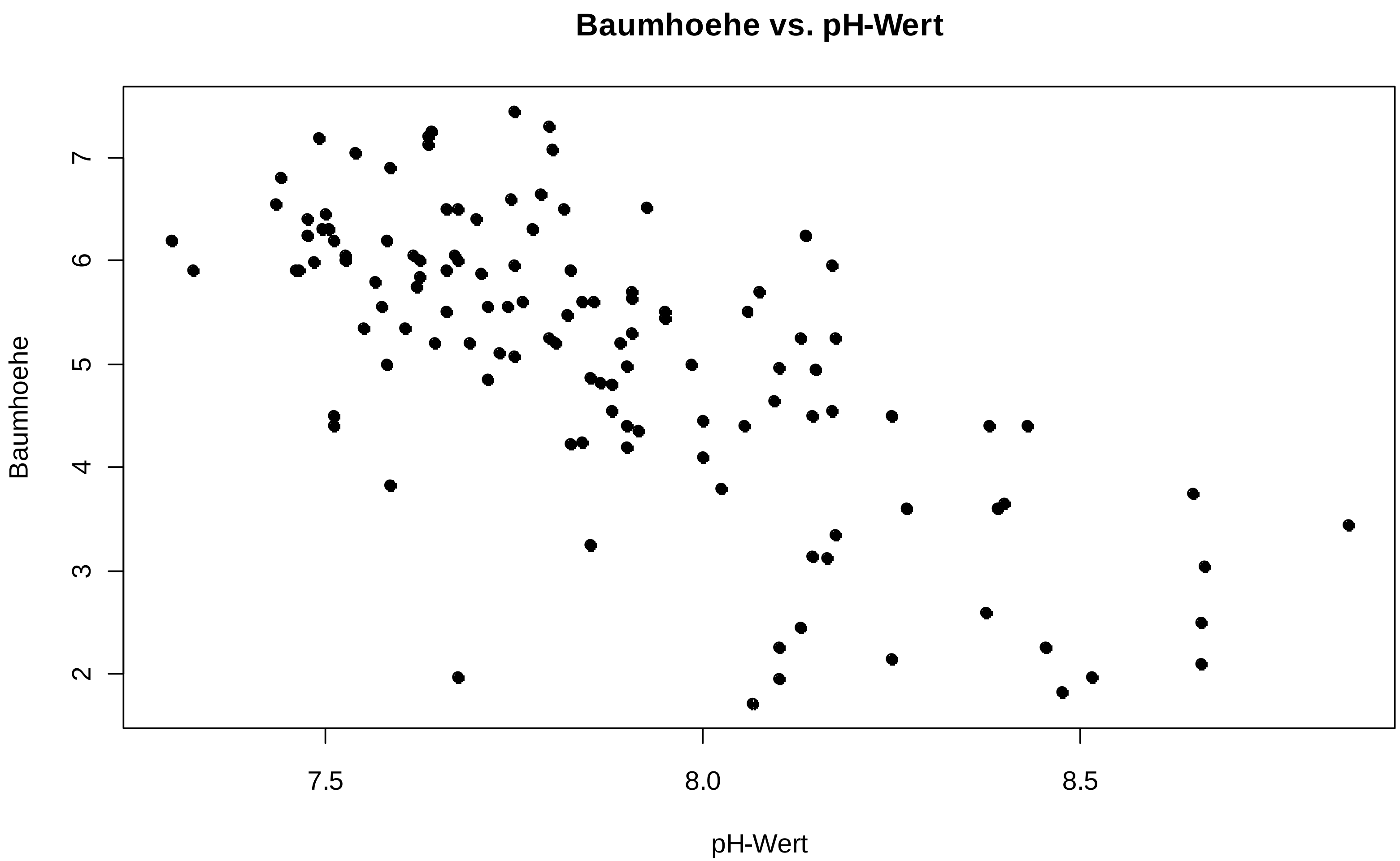
In India, it was observed that alkaline soil hampers plant growth. This gave rise to a search for tree species which show high tolerance against these conditions.

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.

Applied Statistical Regression
HS 2010 – Week 01

Scatterplot: Tree Height vs. pH-value



The Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for all } i=1,\dots,n$$

→ 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!

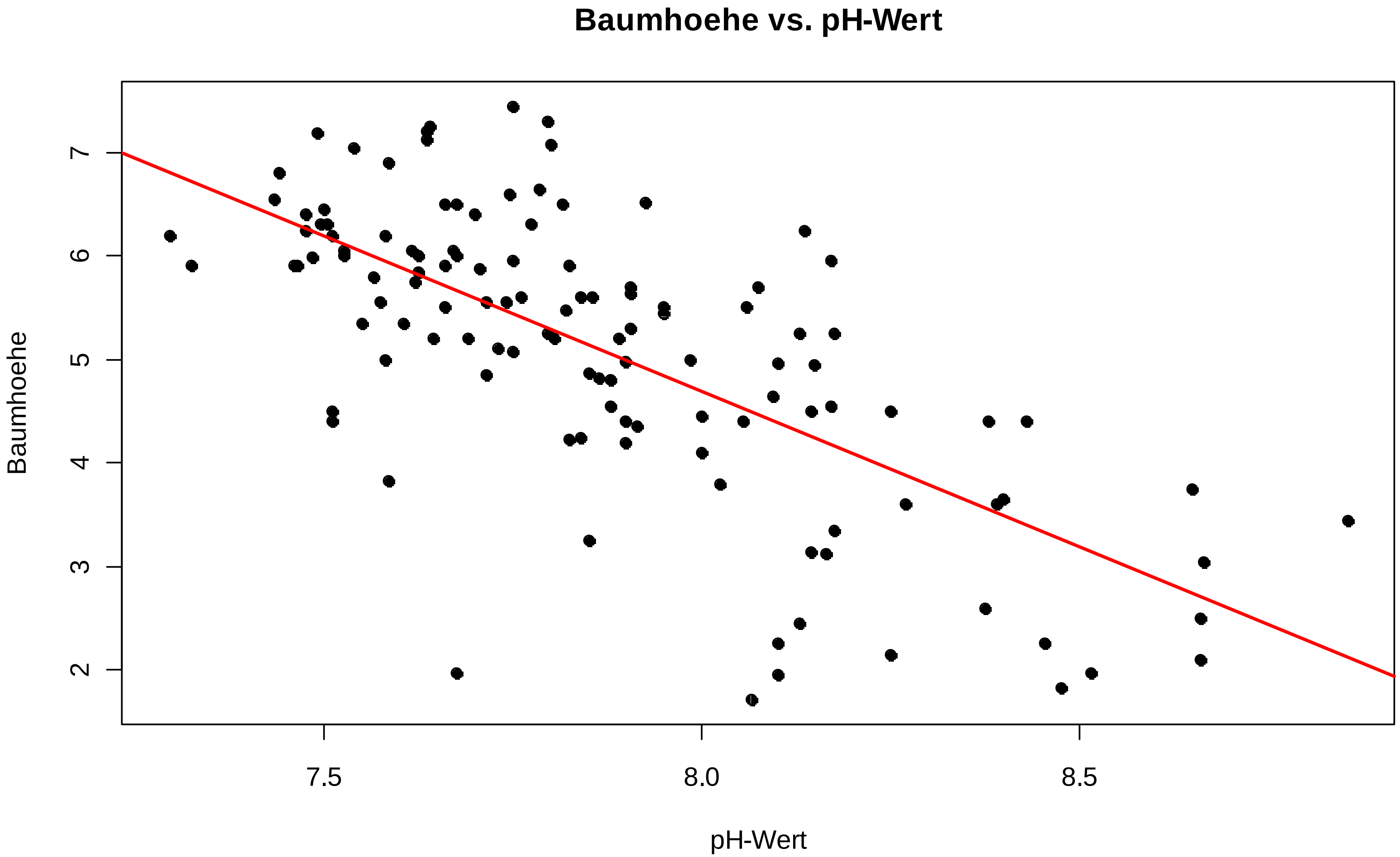
Applied Statistical Regression
HS 2010 – Week 01

Parameter Estimation

→ **See blackboard…**

Applied Statistical Regression
HS 2010 – Week 01

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!

Applied Statistical Regression
HS 2010 – Week 01

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:

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

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

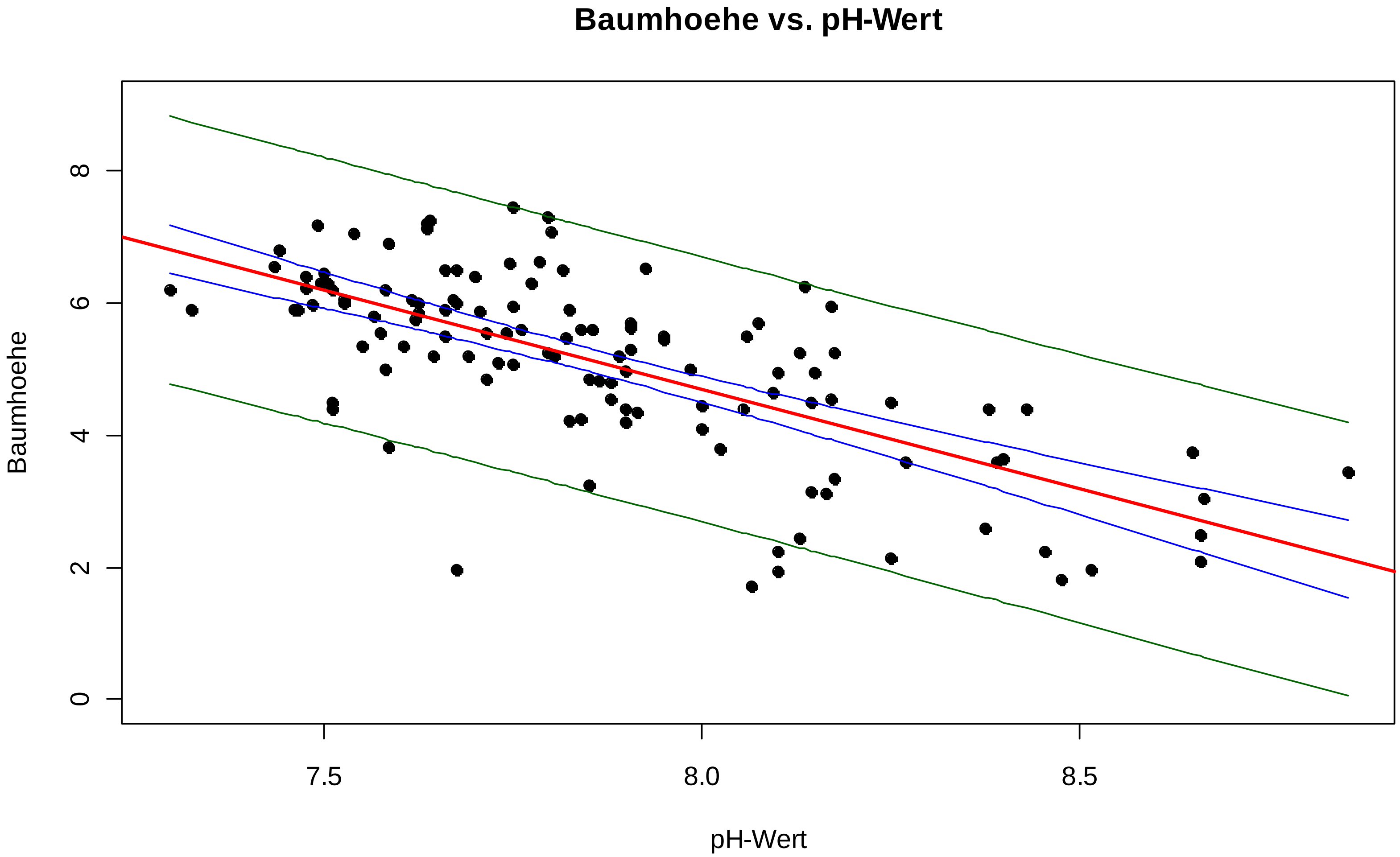
95% confidence interval: this is for the fitted value!

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

95% prediction interval: this is for future observations!

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

Confidence and Prediction Intervals



Applied Statistical Regression
HS 2010 – Week 01

Residual Diagnostics

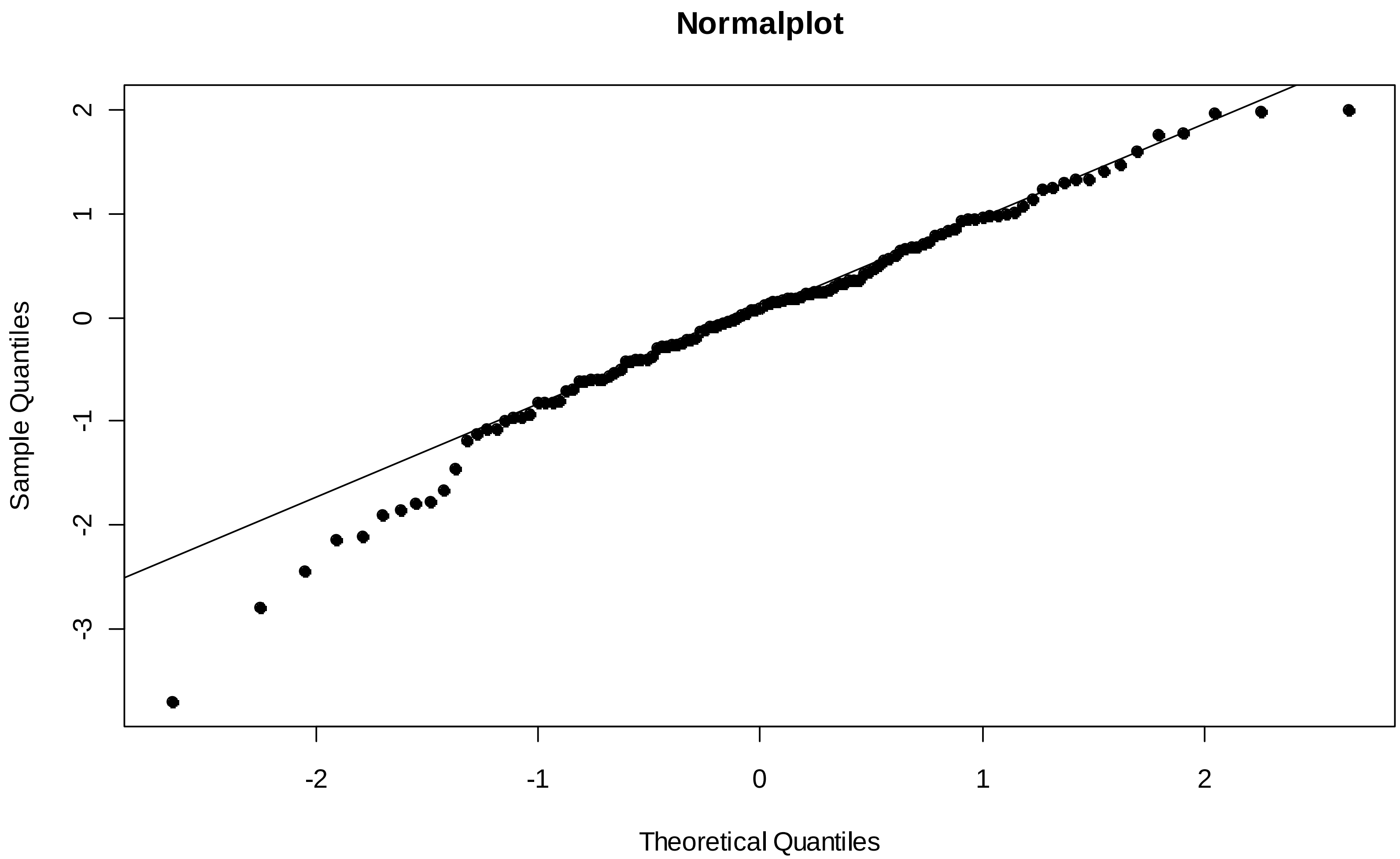
Needs to be done after every regression fit!!!

To check:

- regression line is the correct relation, zero error expected
  → Tukey-Anscombe plot
- scatter is constant, and the residuals are uncorrelated
  → Tukey-Anscombe plot, time series plot
- errors/residuals are normally distributed
  → normal plot

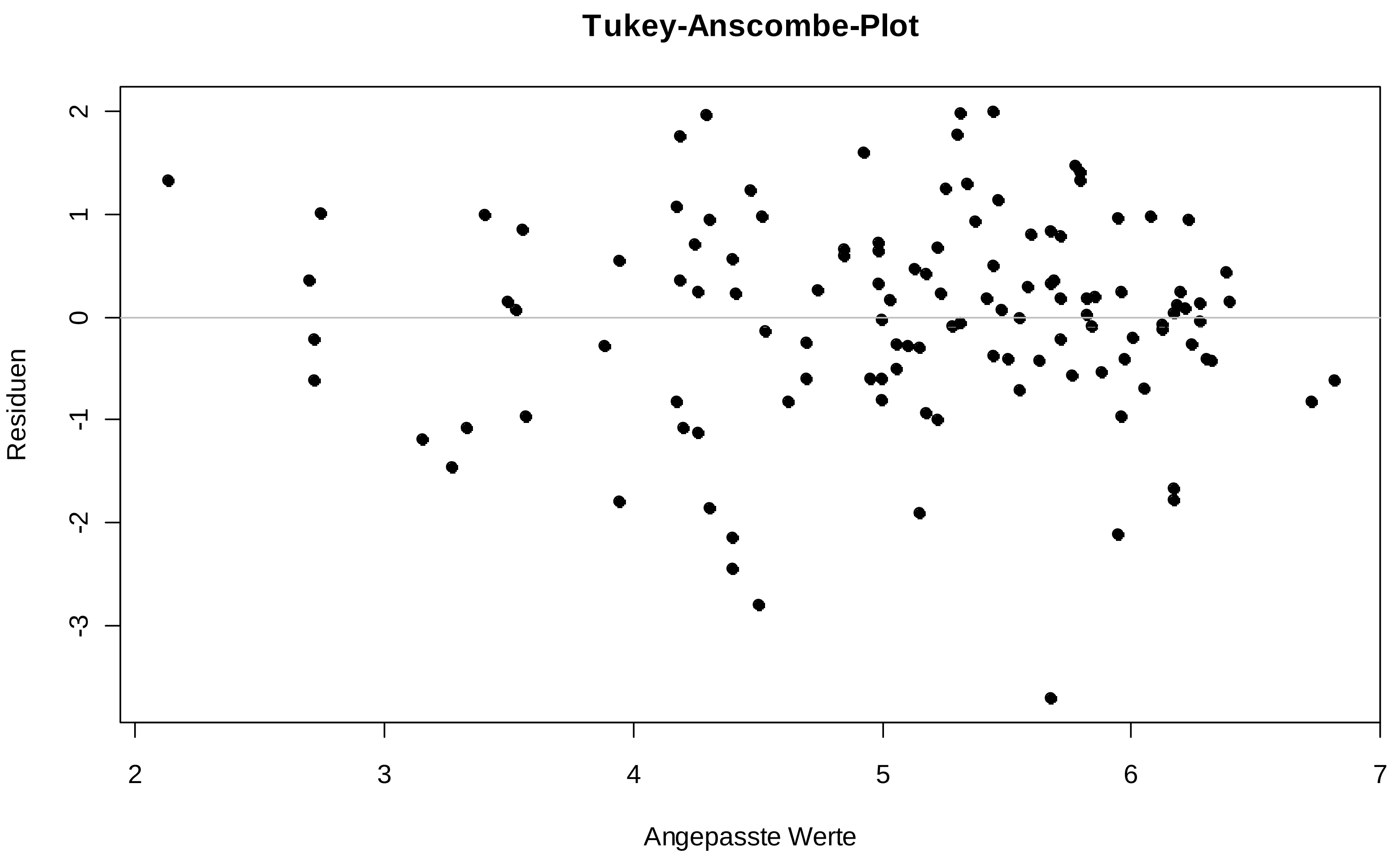
Applied Statistical Regression
HS 2010 – Week 01

Normal Plot



Applied Statistical Regression
HS 2010 – 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

Erroneous Input Variables

What's this?

→ predictors are random, non-deterministic!

→ example: measurement device is not precise

If the usual least squares approach is used, the estimates will be biased:

$$E[\hat{\beta}_1] = \beta_1 \cdot \frac{1}{\left(1+\sigma_\delta^2 / \sigma_\xi^2\right)} \text{, where } \sigma_\xi^2 = \frac{1}{n} \cdot \sum (\xi_i - \overline{\xi})$$

What to do?

→ in case of small errors and prediction only: ignore!

→ for more serious cases, check the work of Draper (1992)