Applied Statistical Regression
HS 2010 – Week 03

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

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

ETH Zürich, October 12, 2010

Course Organization

The exercises will be held on the days that were planned according to the schedule given on the organization sheet!

NEW: the exercise lessons will (until further notice) ALWAYS take place at the computer labs, i.e. in the following rooms:

HG E27	Ag – Go
HG E26.1	Ha – Pa
HG E26.3	Pe – Zh

Applied Statistical Regression
HS 2010 – Week 03

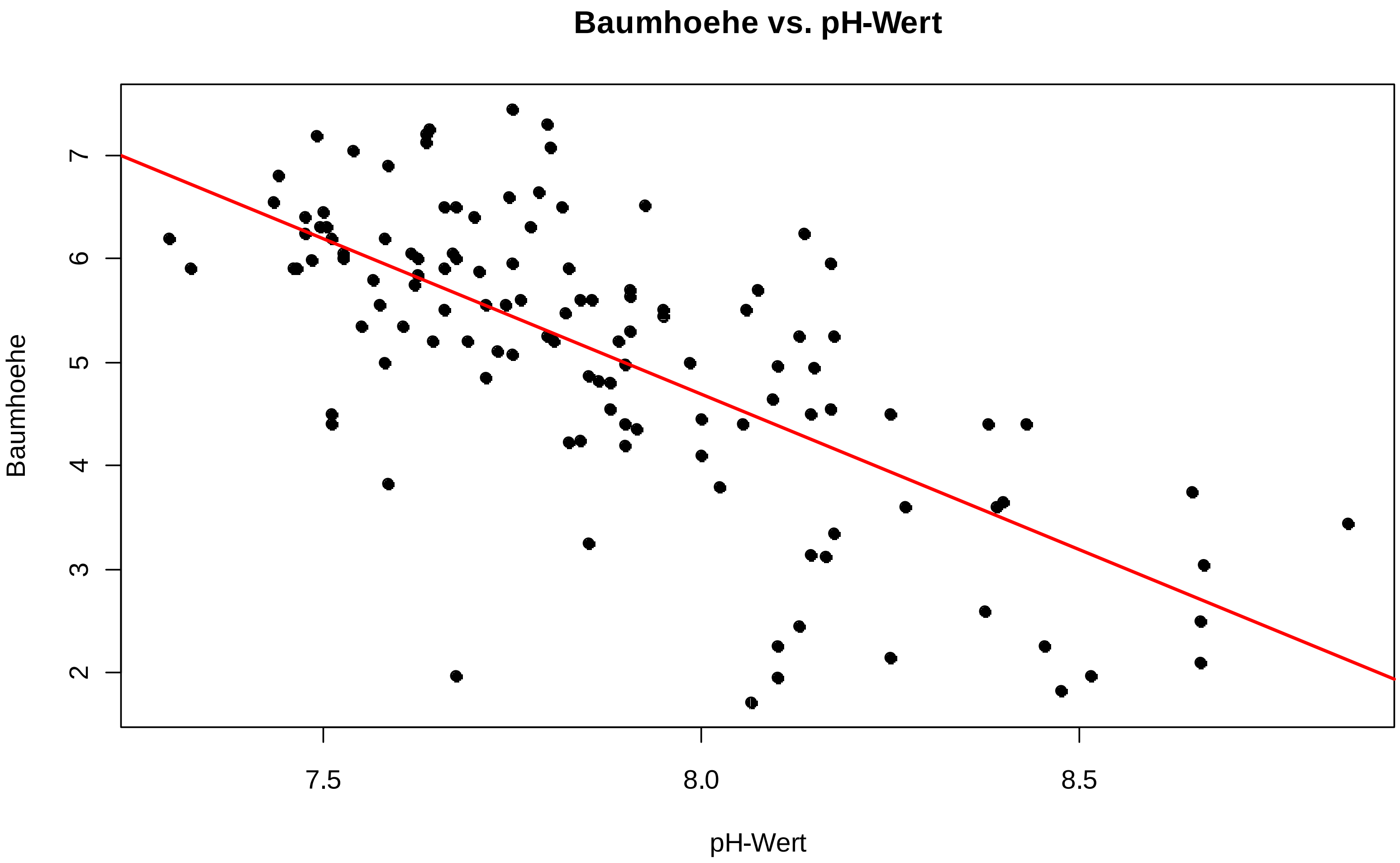
The Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$ 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!

Regression Line



Applied Statistical Regression
HS 2010 – Week 03

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”.

Applied Statistical Regression

HS 2010 – Week 03

Confidence and Prediction Intervals

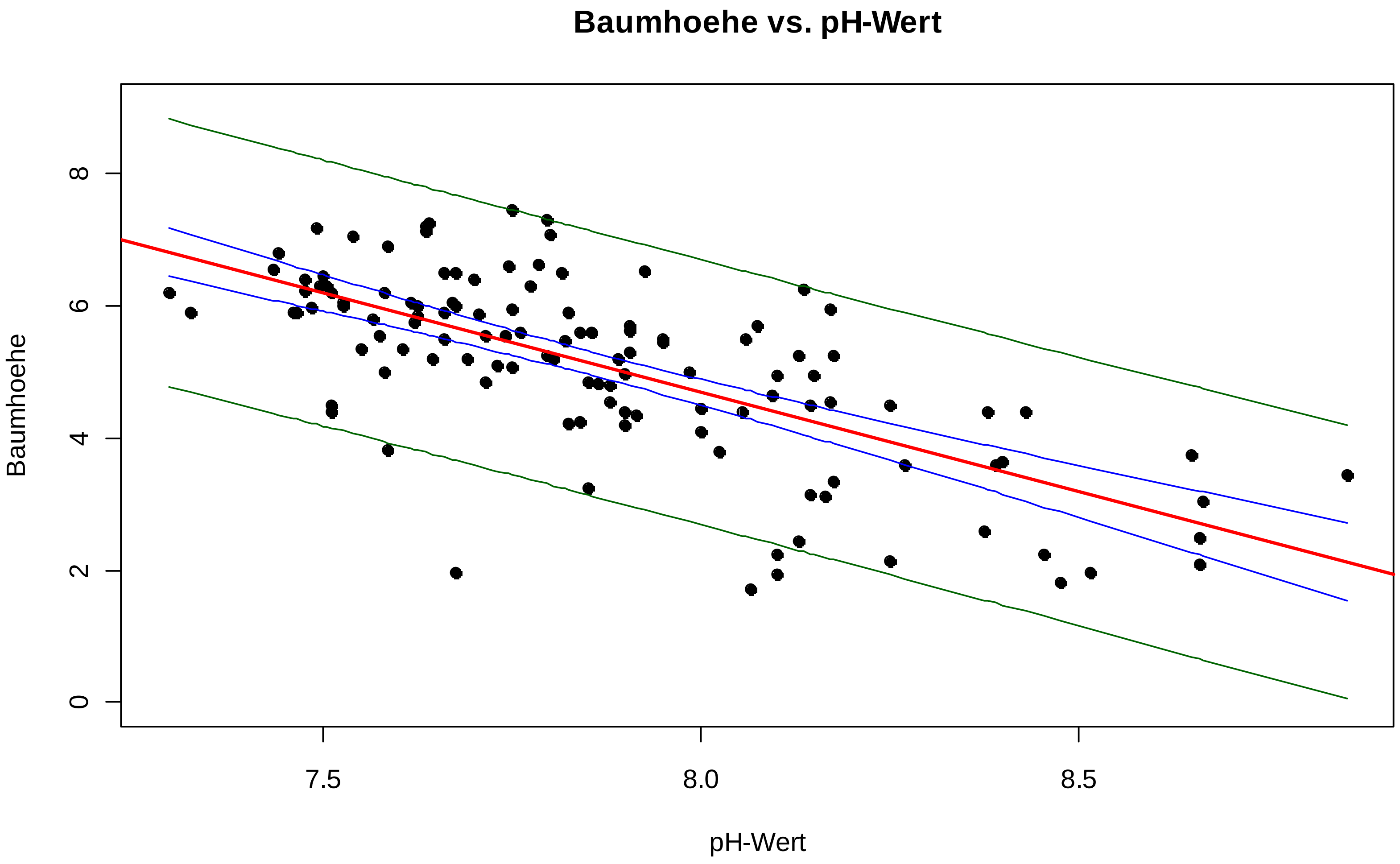
95% confidence interval: this is for the expected value!

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

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



Residual Diagnostics

Needs to be done after every regression fit!!!

To check:

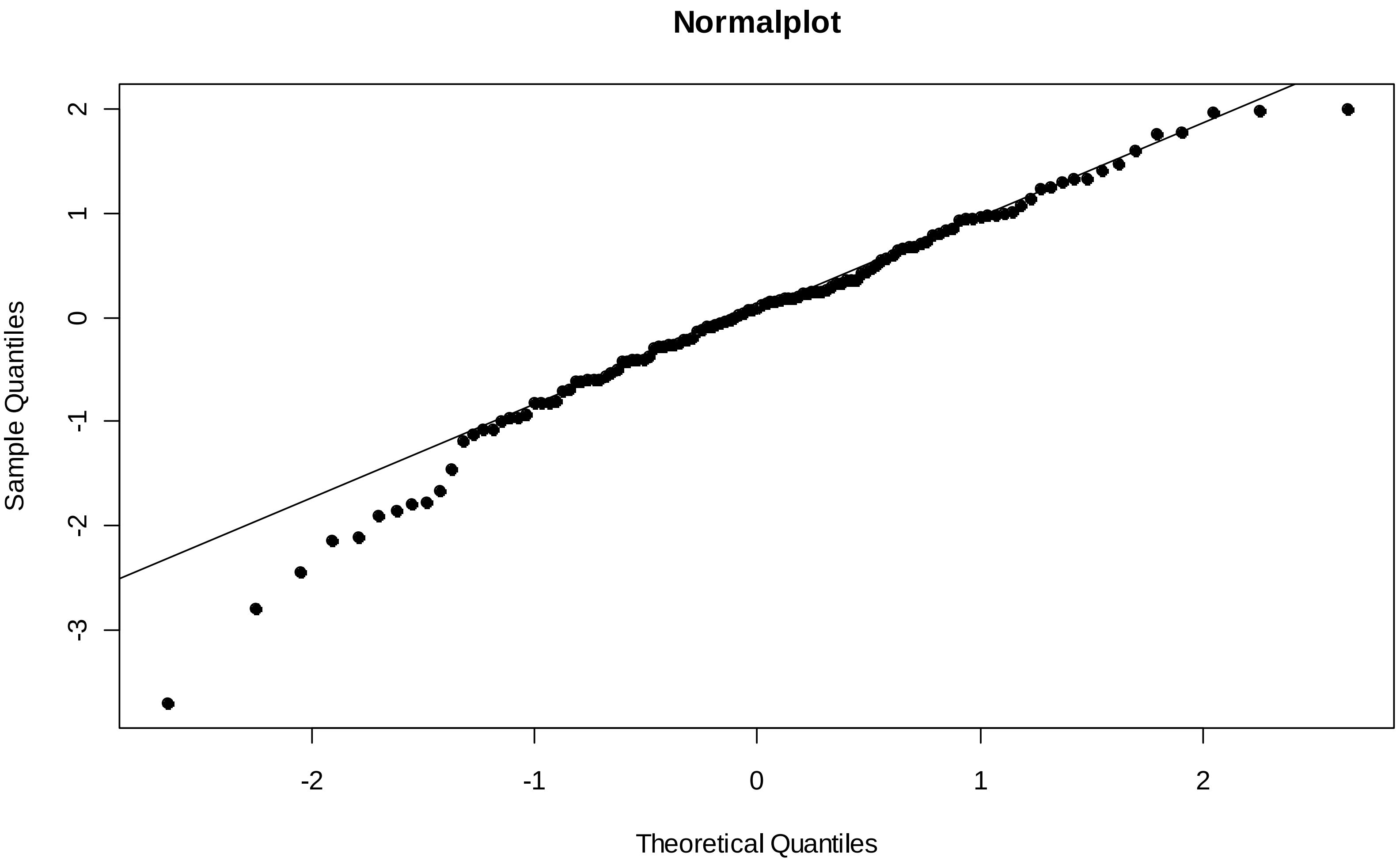
- regression line is the correct relation, zero error expected
  → *Tukey-Anscombe plot*

- scatter is constant, and the residuals are uncorrelated
  → *Tukey-Anscombe plot, time series plot*

- errors/residuals are normally distributed
  → *normal plot*

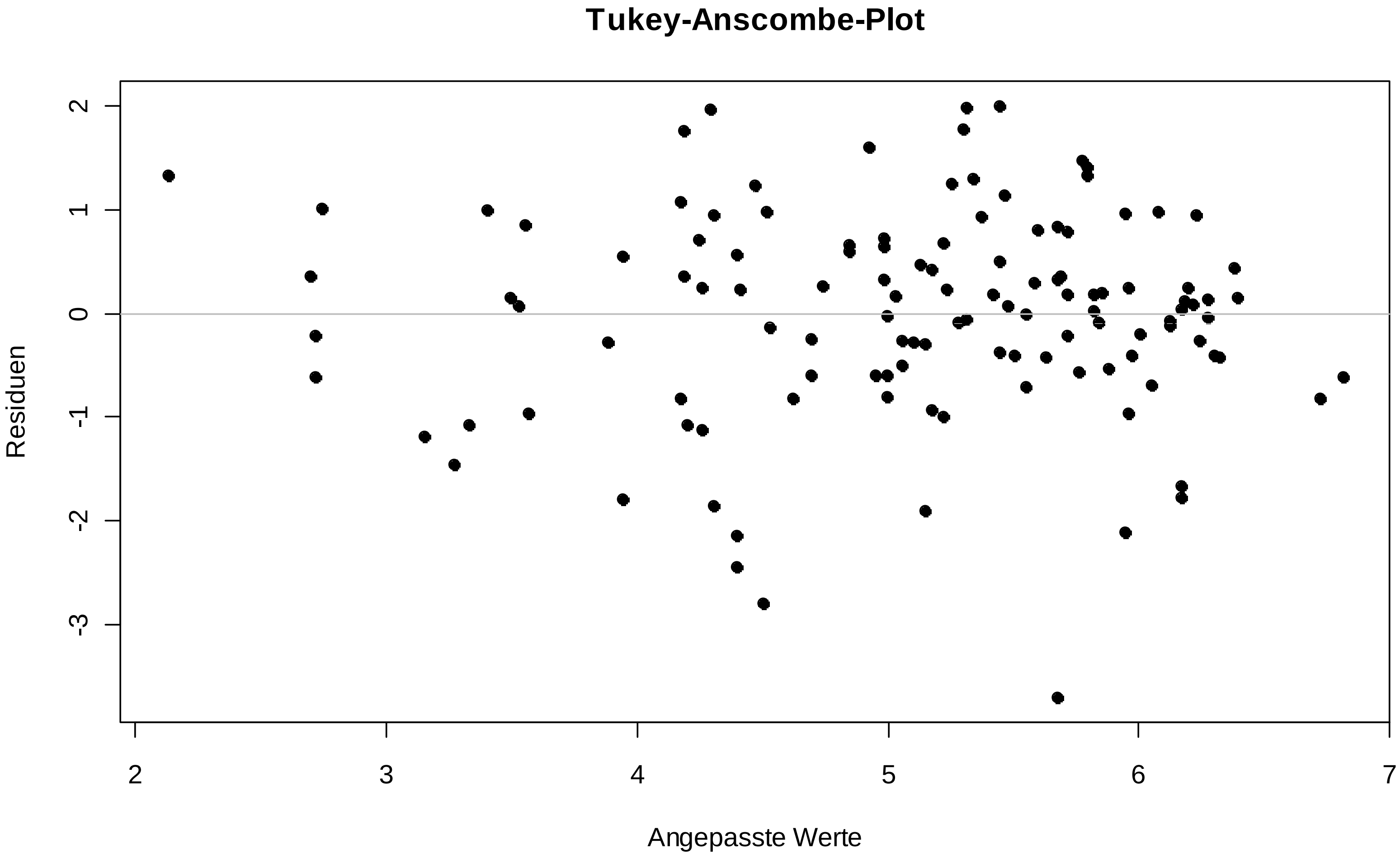
Applied Statistical Regression
HS 2010 – Week 03

Normal Plot



Applied Statistical Regression
HS 2010 – Week 03

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)

Applied Statistical Regression
HS 2010 – Week 03

Multiple Linear Regression

The model is:

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

- we have p predictors now
- visualization is no longer possible
- we are still given n data points, and still:
- the goal is to estimate the regression coefficients

Assumptions on the Error Term

We assumptions are identical to simple linear regression.

- $E[\varepsilon_i] = 0$, i.e. the hyper plane is the correct fit
- $Var(\varepsilon_i) = \sigma_\varepsilon^2$, constant scatter for the error term
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, uncorrelated errors

As in simple linear regression, we do not require any specific distribution for parameter estimation and certain optimality results of the least squares approach. The distributional assumption only comes into play when we do inference on the parameters.

Don‘t Do Many Simple Regressions

Doing many simple linear regressions is not equivalent to multiple linear regression. Check the example

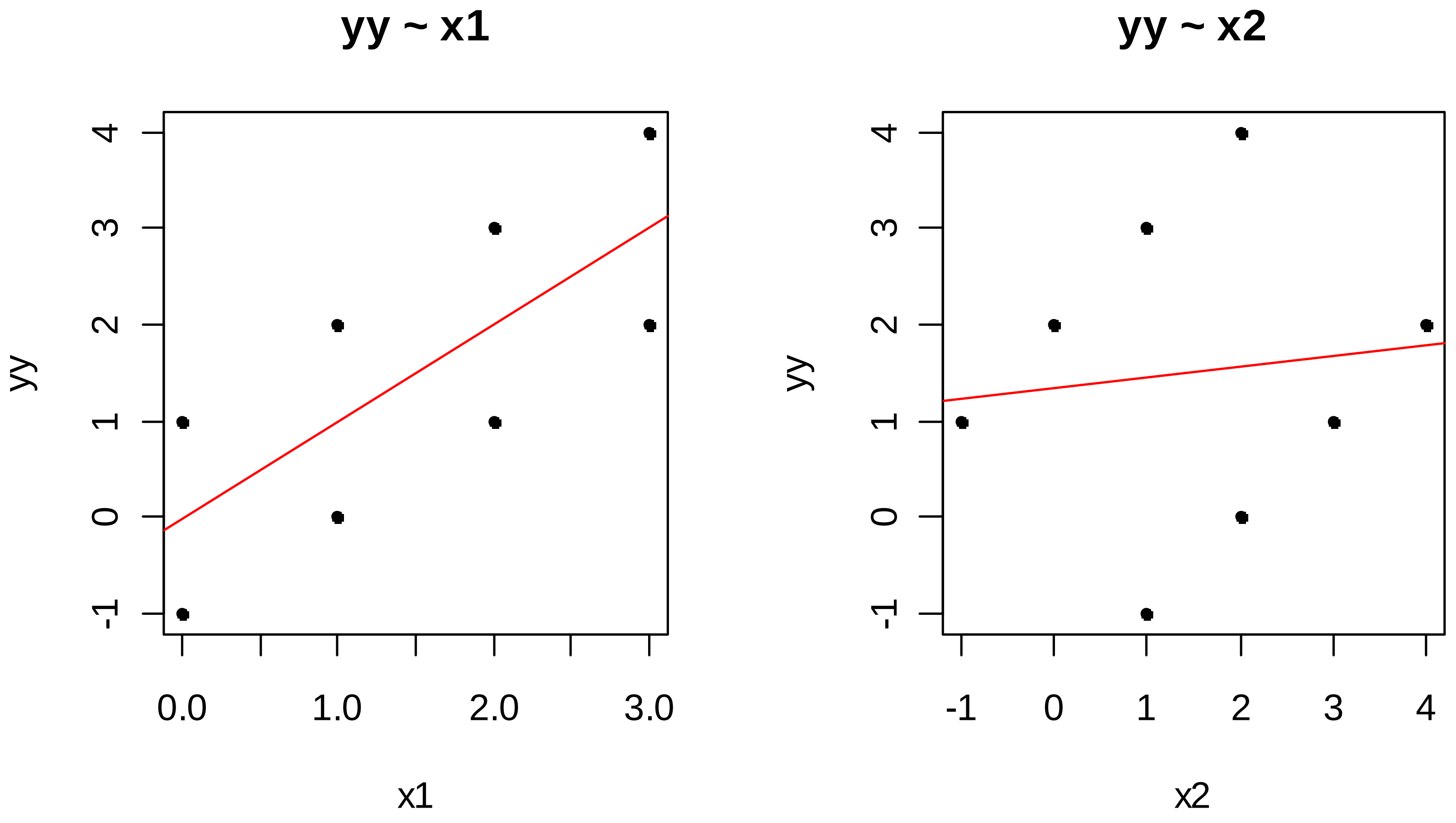
x1	0	1	2	3	0	1	2	3
x2	-1	0	1	2	1	2	3	4
yy	1	2	3	4	-1	0	1	2

We have $Y_i = \hat{y}_i = 2x_{i1} - x_{i2}$, a perfect fit.

Thus, all residuals are 0 and $\hat{\sigma}_{\varepsilon}^2$.

→ *But what is the result from simple linear regressions?*

Don‘t Do Many Simple Regressions



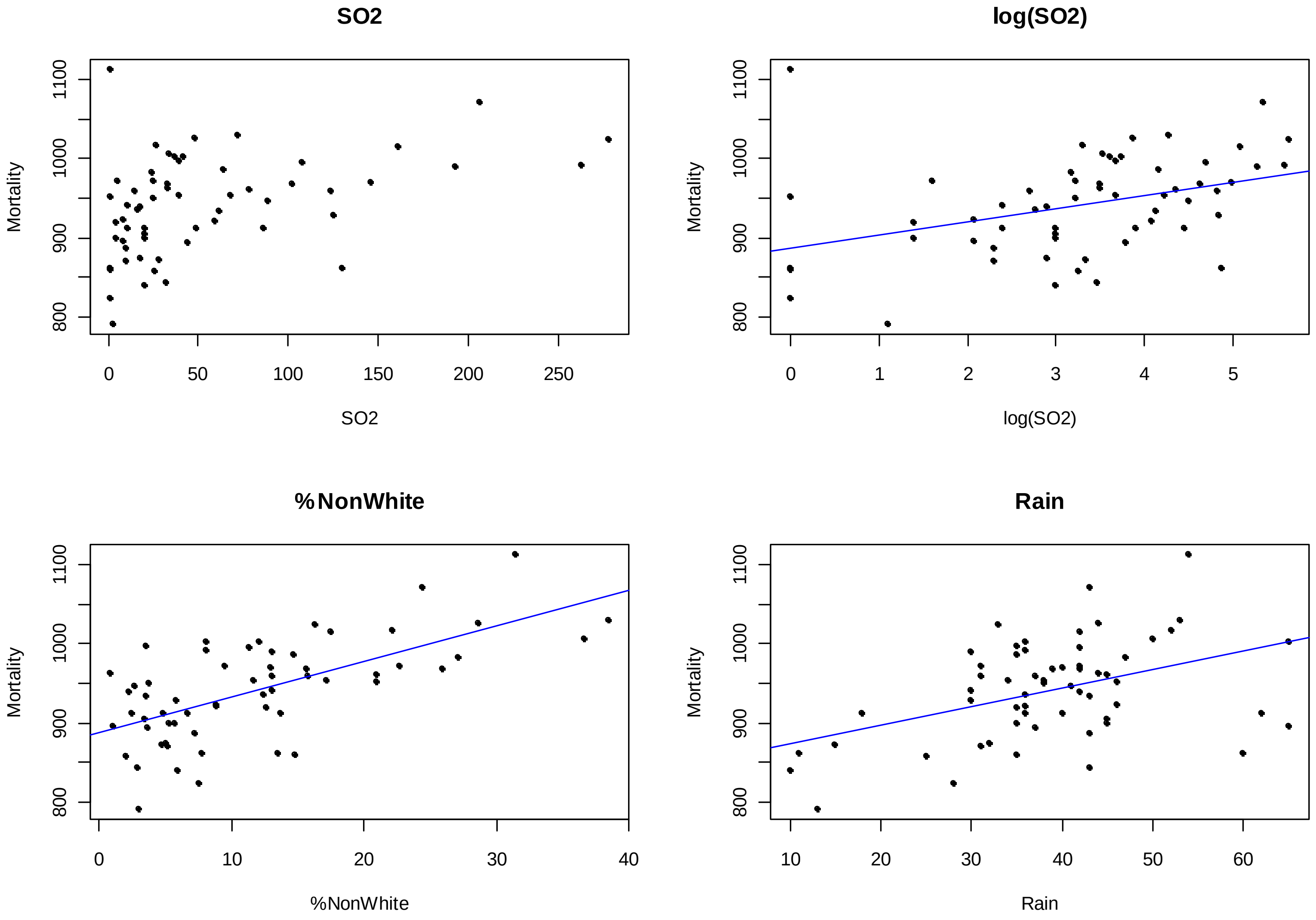
Applied Statistical Regression
HS 2010 – Week 03

An Example

City	Mortality	JanTemp	JulyTemp	RelHum	Rain	Educ	Dens	NonWhite	WhiteCollar	Pop	House	Income	HC	NOx	SO2
Akron, OH	921.87	27	71	59	36	11.4	3243	8.8	42.6	660328	3.34	29560	21	15	59
Albany, NY	997.87	23	72	57	35	11	4281	3.5	50.7	835880	3.14	31458	8	10	39
Allentown, PA	962.35	29	74	54	44	9.8	4260	0.8	39.4	635481	3.21	31856	6	6	33
Atlanta, GA	982.29	45	79	56	47	11.1	3125	27.1	50.2	2138231	3.41	32452	18	8	24
Baltimore, MD	1071.29	35	77	55	43	9.6	6441	24.4	43.7	2199531	3.44	32368	43	38	206
Birmingham, AL	1030.38	45	80	54	53	10.2	3325	38.5	43.1	883946	3.45	27835	30	32	72

Applied Statistical Regression
HS 2010 – Week 03

Some Simple Linear Regressions



Coefficient Estimates

log(SO2): $\hat{y} = 886.34 + 16.86 \cdot \log(SO_2)$

NonWhite: $\hat{y} = 887.90 + 4.49 \cdot NonWhite$

Rain: $\hat{y} = 851.22 + 2.34 \cdot Rain$

```
> lm(Mortality ~ log(SO2) + NonWhite + Rain, data=mortality)
> Coefficients:
> (Intercept)    log(SO2)    NonWhite        Rain
    773.020      17.502      3.649           1.763
```

The regression coefficient is the increase in the response, if the predictor increases by 1 unit, but all other predictors remain unchanged.

Least Squares Approach

We determine residuals

$$r_i = y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})$$

Then, we choose the parameters such that the sum of squared residuals

$$\sum_{i=1}^{n} r_i^2$$

is minimal. As in simple linear regression, there is an explicit solution to this problem. It can be attained by taking partial derivatives and setting them to zero. This again results in the so-called *normal equations.*

Applied Statistical Regression
HS 2010 – Week 03

Matrix Notation

In matrix notation, the multiple linear regression model can be written as:

$$Y = X\beta + \varepsilon$$

The elements in this equation are as follows:

→ see blackboard…

Applied Statistical Regression
HS 2010 – Week 03

Normal Equations and Their Solutions

The least squares approach leads to the normal equations, which are of the following form:

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

- Unique solution if and only if X has full rank
- Predictor variables need to be linearly independent

- If X has not full rank, the model is “badly formulated”
- Design improvement mandatory!!!

- Necessary (not sufficient) condition: p<n
- Do not over-parametrize your regression!

Properties of the Estimates

Gauss-Markov-Theorem:

The regression coefficients are unbiased estimates, and they fulfill the optimality condition of minimal variance among all linear, unbiased estimators (*BLUE*).

- $E[\hat{\beta}] = \beta$
- $Cov(\beta) = \sigma_{\varepsilon}^{2} \cdot (X^{T}X)^{-1}$
- $\hat{\sigma}_{\varepsilon}^{2} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} r_i^2$ (note: degrees of freedom!)

Hat Matrix Notation

The fitted values are:

$$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = HY$$

The matrix is called hat matrix, because “it puts a hat on the Y’s”, i.e. transforms the observed values into fitted values. We can also use this matrix for computing the residuals:

$$r = Y - \hat{Y} = (I - H)Y$$

Moments of these estimates:

$$E[\hat{y}] = y\,,\; E[r] = 0$$

$$Var(\hat{y}) = \sigma_\varepsilon^2 H\;,\; Var(r) = \sigma_\varepsilon^2 (I - H)$$

If the Errors are Gaussian…

While all of the above statements hold for arbitrary error distribution, we obtain some more, very useful properties by assuming i.i.d. Gaussian errors:

- $\hat{\beta} \sim N\left(\beta, \sigma_\varepsilon^2 (X^T X)^{-1}\right)$
- $\hat{y} \sim N(X\beta, \sigma_\varepsilon^2 H)$
- $\hat{\sigma}_\varepsilon^2 \sim \frac{\sigma_\varepsilon^2}{n-p} \chi_{n-p}$

What to do if the errors are non-Gaussian?

Applied Statistical Regression
HS 2010 – Week 03

Individual Parameter Tests

If we are interested whether the jth predictor variable is relevant, we can test the hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_A : \beta_j \neq 0$$

We can derive the test statistic and its distribution:

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_\varepsilon^2 (X^T X)_{jj}^{-1}}} \sim t_{n-(p+1)}$$

Applied Statistical Regression
HS 2010 – Week 03

Individual Parameter Tests

These tests quantify the effect of the predictor x_j on the response Y after having subtracted the linear effect of all other predictor variables on Y.

Be careful, because of:

a) The *multiple testing problem*: when doing many tests, the total type II error increases. By how much: see blackboard

b) It can happen that all individual tests do not reject the null hypothesis, although some predictors have a significant effect on the response. Reason: correlated predictors!

Global F-Test

Question: is there *any* relation between predictors and response?

We test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

against the alternative

$H_A : \beta_j \neq 0$ for at least one j in 1,…, p

The test statistic is:

$$F = \frac{n-(p+1)}{p} \cdot \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \sim F_{p,n-(p+1)}$$

Applied Statistical Regression
HS 2010 – Week 03

Coefficient of Determination

The coefficient of determination, also called *multiple R-squared*, is aimed at describing the goodness-of-fit of the multiple linear regression model:

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

It shows the proportion of the total variance which has been explained by the predictors. The extreme cases 0 and 1 mean:…

Applied Statistical Regression
HS 2010 – Week 03

Adjusted Coefficient of Determination

If we add more and more predictor variables to the model, R-squared will always increase, and never decreases

Is that a realistic goodness-of-fit measure?
→ **NO, we better adjust for the number of predictors!**

$$adjR^2 = 1 - \frac{n-1}{n-(p+1)} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \in [0,1]$$

Applied Statistical Regression
HS 2010 – Week 03

R-Output

```
> summary(lm(Mortality~log(SO2)+NonWhite+Rain, data=mo…))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 773.0197    22.1852  34.844  < 2e-16 ***
log(SO2)     17.5019     3.5255   4.964 7.03e-06 ***
NonWhite      3.6493     0.5910   6.175 8.38e-08 ***
Rain          1.7635     0.4628   3.811 0.000352 ***
---
Residual standard error: 38.4 on 55 degrees of freedom
Multiple R-squared: 0.641,  Adjusted R-squared: 0.6214
F-statistic: 32.73 on 3 and 55 DF,  p-value: 2.834e-12
```

Interpreting the Result

Does the SO2 concentration affect the mortality?

→ Might be, might not be

→ There are only 3 predictors

→ We could suffer from confounding effects

→ Causality is always difficult, but…

The next step is to include all predictor variables that are present in the mortality dataset.

Applied Statistical Regression

HS 2010 – Week 03

More Predictors

```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.164e+03  2.939e+02   3.960 0.000258 ***
JanTemp     -1.669e+00  7.930e-01  -2.105 0.040790 *  
JulyTemp    -1.167e+00  1.939e+00  -0.602 0.550207    
RelHum       7.017e-01  1.105e+00   0.635 0.528644    
Rain         1.224e+00  5.490e-01   2.229 0.030742 *  
Educ        -1.108e+01  9.449e+00  -1.173 0.246981    
Dens         5.623e-03  4.482e-03   1.255 0.215940    
NonWhite     5.080e+00  1.012e+00   5.019 8.25e-06 ***
WhiteCollar -1.925e+00  1.264e+00  -1.523 0.134623    
Pop          2.071e-06  4.053e-06   0.511 0.611799    
House       -2.216e+01  4.040e+01  -0.548 0.586074    
Income       2.430e-04  1.328e-03   0.183 0.855617    
log(SO2)     6.833e+00  5.426e+00   1.259 0.214262    
---
Residual standard error:  36.2 on 46 degrees of freedom
Multiple R-squared: 0.7333,  Adjusted R-squared: 0.6637 
F-statistic: 10.54 on 12 and 46 DF,  p-value: 1.417e-09
```

Applied Statistical Regression
HS 2010 – Week 03

Some Thoughts on Collinearity

a) With collinear predictors, inference (i.e. interpreting p-values from individual parameter tests and the global F-test) should be "handled with care"!

b) Drawing conclusions on causality should be left out.

c) However, the fitted values are not affected by this, and also prediction with a model fitted from collinear predictors is always fine.

Measuring collinearity: $VIF_j = \dfrac{1}{1-R_j^2}$