Zurich University
of Applied Sciences

Applied Statistical Regression
HS 2011 – Week 04

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 17, 2011

Curvilinear Fitting

All models such as:

- $Y_i = \beta_0 + \beta_1 \cdot \ln(x_i) + E_i$
- $Y_i = \beta_0 + \beta_1 \cdot \sqrt{x} + E_i$
- $Y_i = \beta_0 + \beta_1 \cdot x^{-1} + E_i$

Are simple linear regression models. There is only one single predictor, and the relation is linear in the parameters.

None of these models fits a straight line in the scatterplot, these are all curvilinear relations – linear regression is very versatile!

Applied Statistical Regression
HS 2011 – Week 04

Logged Predictor and Response

Regression models of the form

$$Y_i' = \beta_0' + \beta_1 \cdot x_i' + E_i'$$

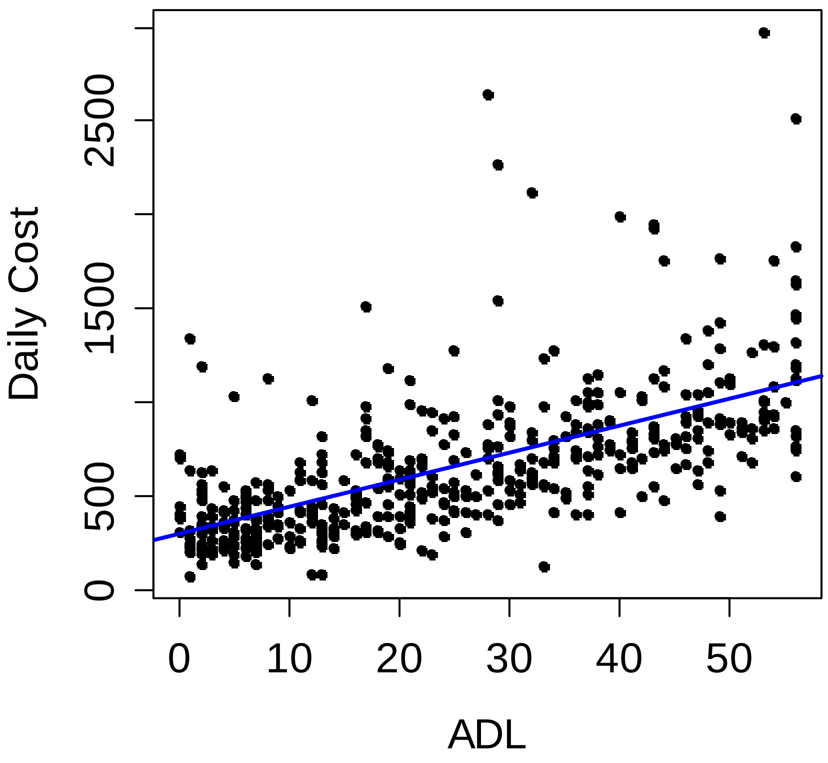
where $Y_i' = \log(Y_i)$ and $x_i' = \log(x_i)$ are very important and often encountered in practice. Backtransformation shows that the initial relation is:

$$Y_i = \beta_0 \cdot x_i^{\beta_1} \cdot E_i$$

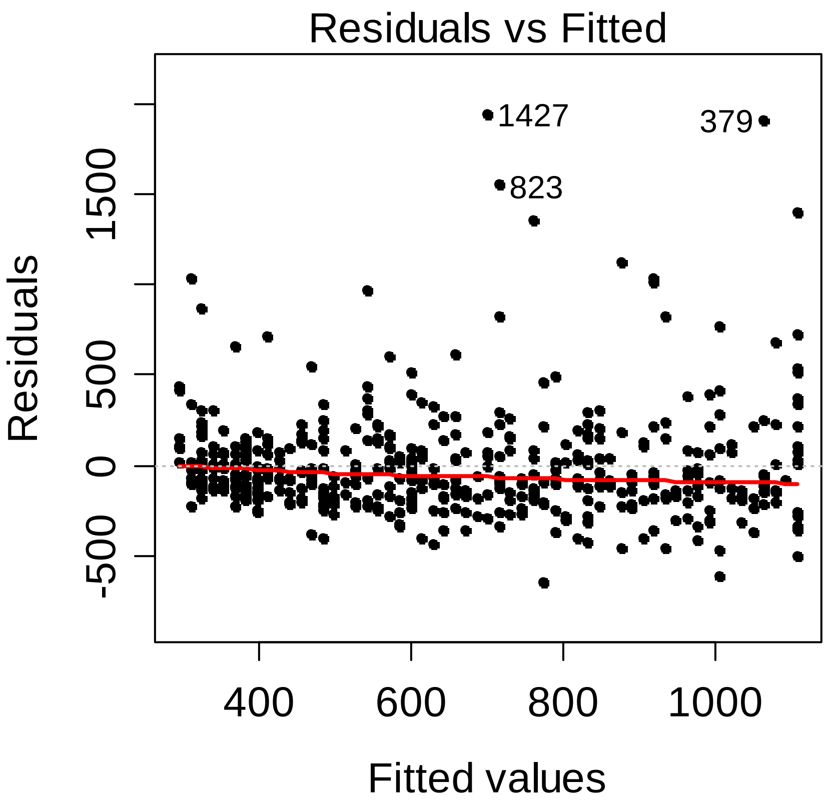
i.e. a non-linear relation with multiplicative error. Through the transformation, the parameter estimation problem is linearized, and can be solved with the least squares method.

Example: Daily Cost in Rehabilitation

Daily Cost in Rehab vs. ADL



Residuals vs. Fitted Values



Logged Response Model

We transform the response variable and try to explain it using a linear model with our previous predictors:

$$Y' = \log(Y) = \beta_0 + \beta_1 x + E$$

In the original scale, we can write the logged response model using the same predictors:

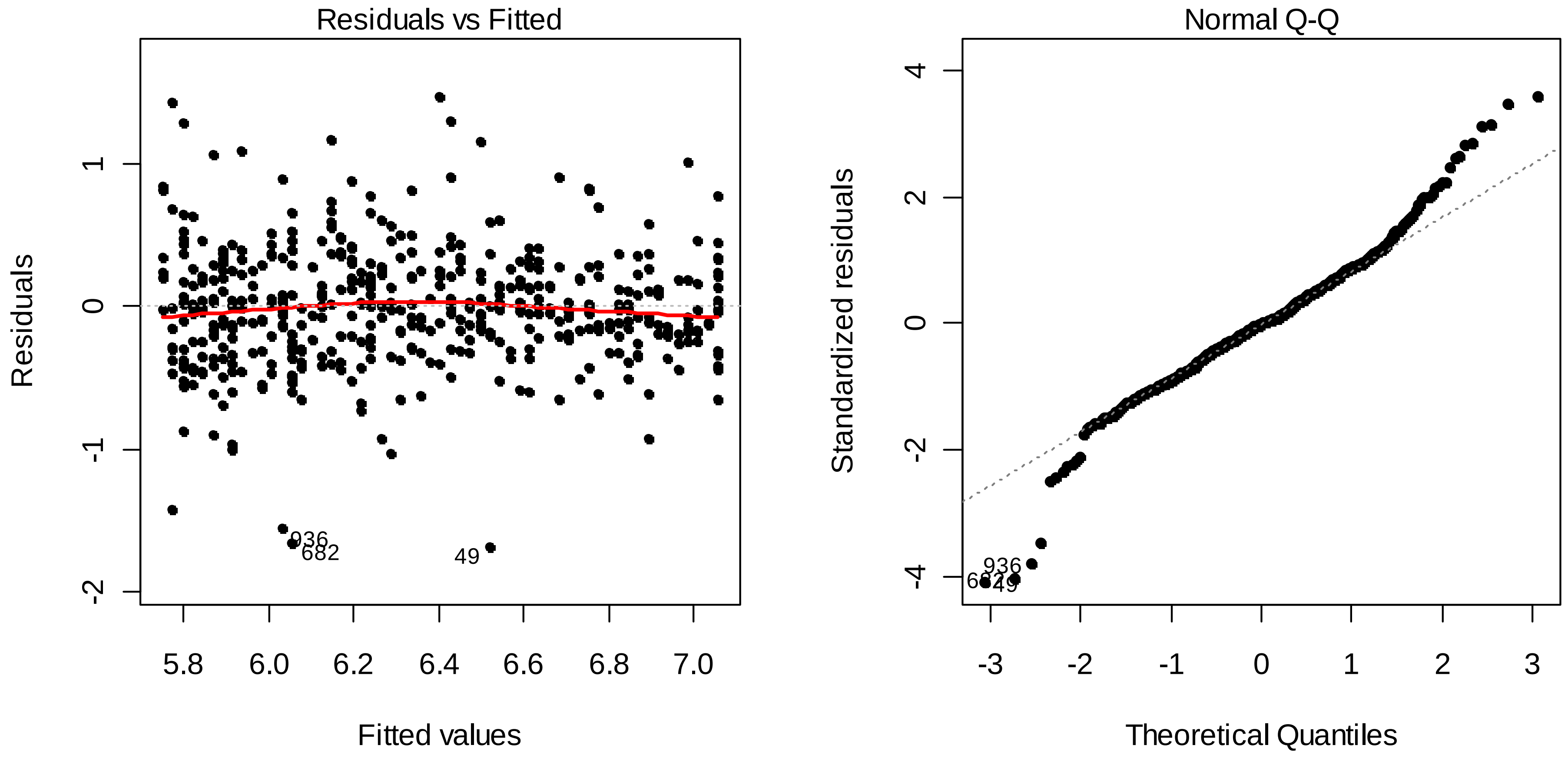
$$Y = \exp(\beta_0 + \beta_1 x) \cdot \exp(E)$$

→ Multiplicative model

→ $E \sim N(0, \sigma_E^2)$, and thus, $\exp(E)$ has a lognormal distribution

Applied Statistical Regression
HS 2011 – Week 04

Also This Transformation Works!



Dealing with Zero Response

- Logged response model is only applicable when the response is strictly positive…
- What if there are some cases with $Y = 0$?
  - never omit these
  - additive shifting is possible
- How to additively shift?
  - usual choice: c=1
  - not good, because effect is scale-dependent

→ **Shift with the value of the smallest positive observation!**

Applied Statistical Regression

HS 2011 – Week 04

Back Transforming the Fitted Values

- In principle, we can „simply back transform“

$$\hat{y} = \exp(\hat{y}')$$

- This is an estimate for the median, but not the mean!

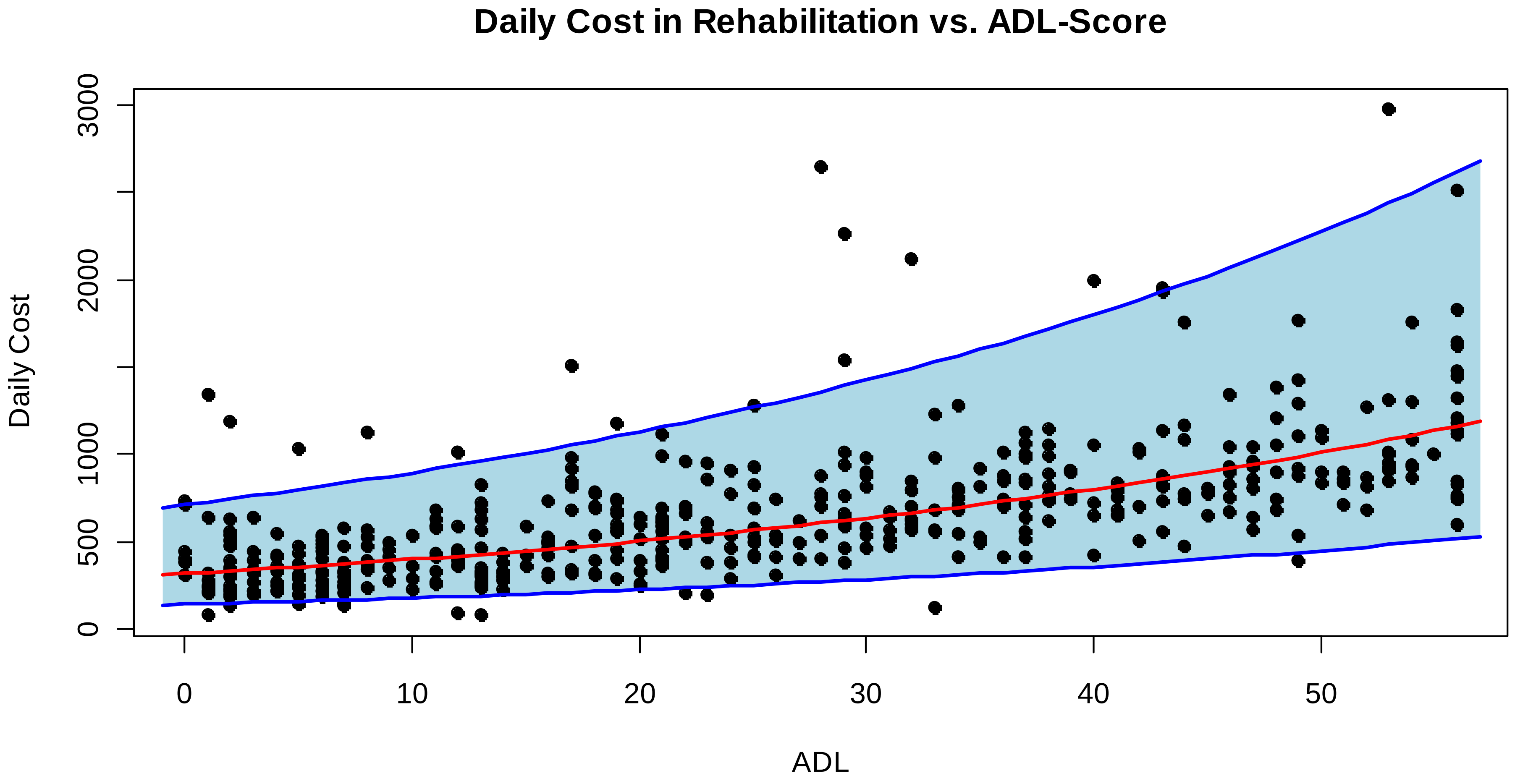
- If unbiased estimation is required, then use:

$$\hat{y} = \exp\left(\hat{y}' + \frac{\hat{\sigma}_E^2}{2}\right)$$

- Confidence/prediction intervals are not problematic

$$[l,u] \;\rightarrow\; [\exp(l), \exp(u)]$$

Back Transforming: Example



Interpretation of the Coefficients

Important: there is no back transformation for the coefficients to the original scale, but still a good interpretation

$$\begin{aligned}\log(\hat{y}) &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p \\ \hat{y} &= \exp(\hat{\beta}_0)\exp(\hat{\beta}_1 x_1)\ldots\exp(\hat{\beta}_p x_p)\end{aligned}$$

An increase by one unit in x_1 would multiply the fitted value in the original scale with $\exp(\hat{\beta}_1)$.

→ **Coefficients are interpreted multiplicatively!**

First-Aid Transformations

These are intendend to stabilize the variance

First-Aid Transformations:

→ do always apply these (if no practical reasons against it)

→ to both response and predictors

Absolute values and concentrations:

log-transformation: $y' = \log(y)$

Count data:

square-root transformation: $y' = \sqrt{y}$

Proportions:

arcsine transformation: $y' = \sin^{-1}\left(\sqrt{y}\right)$

Applied Statistical Regression
HS 2011 – Week 04

Multiple Linear Regression

The model is:

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

- 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.

- $E[E_i]=0$, i.e. the hyper plane is the correct fit
- $Var(E_i)=\sigma_E^2$, constant scatter for the error term
- $Cov(E_i,E_j)=0$, 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

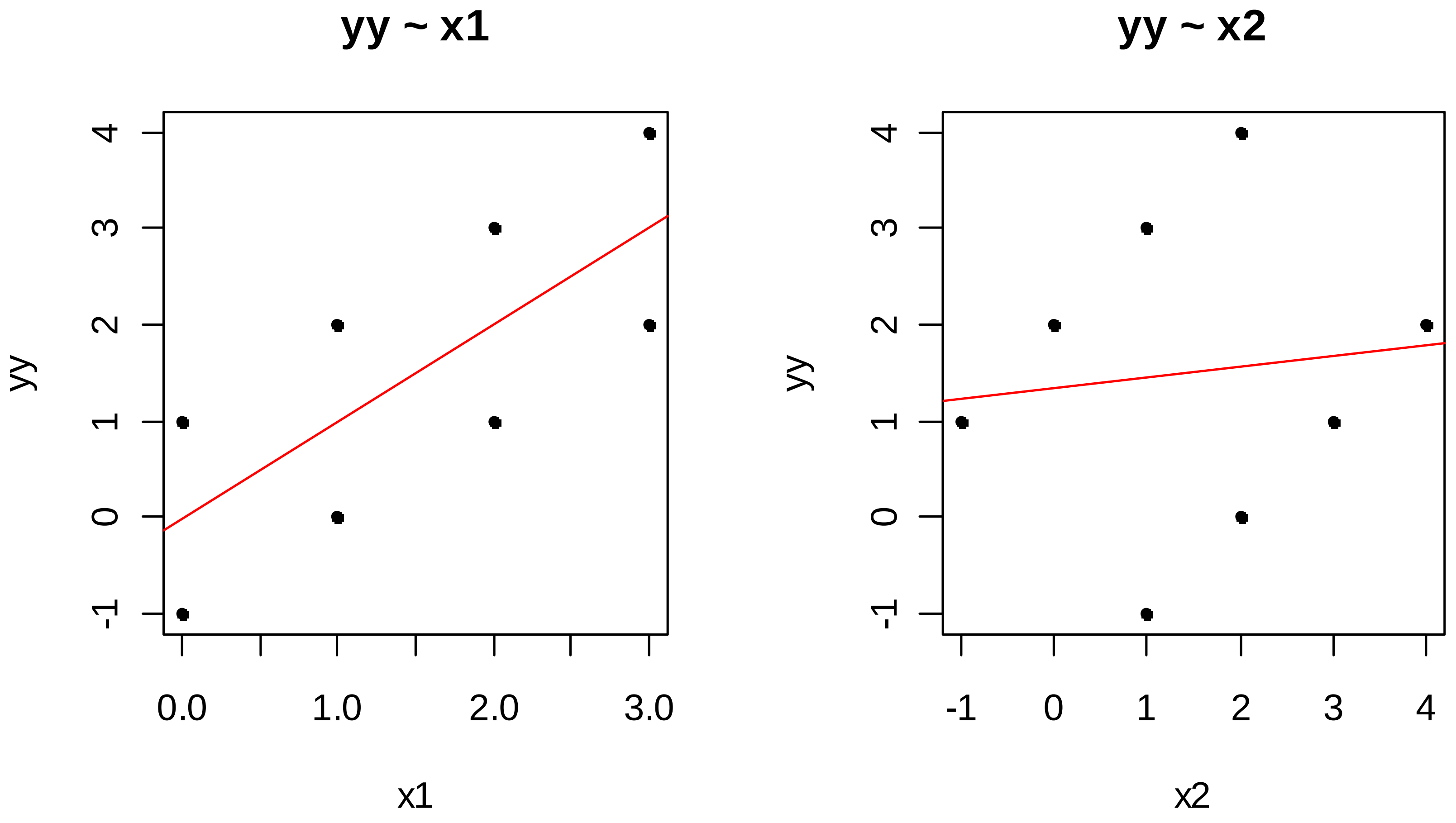
We have $Y_i = \hat{y}_i = 2x_{i1} - x_{i2}$, a perfect fit.

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

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

Applied Statistical Regression
HS 2011 – Week 04

Don‘t Do Many Simple Regressions



An Example

Researchers at General Motors collected data on 60 US Standard Metropolitan Statistical Areas (SMSAs) in a study of whether air pollution contributes to mortality.

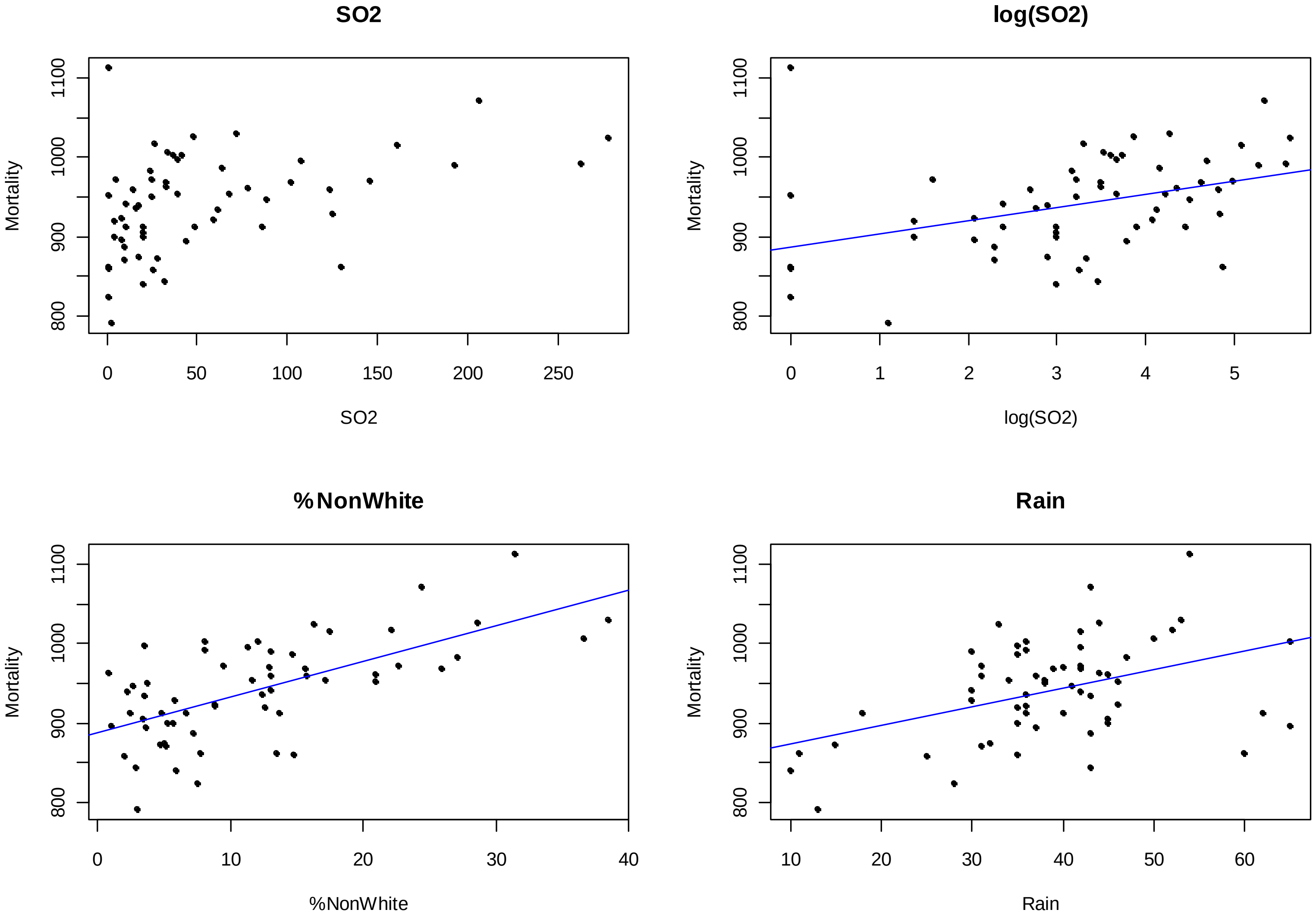
City	Mortality	JanTemp	JulyTemp	RelHum	Rain	Educ	Dens	NonWhite	WhiteCollar	Pop	House	Income	HC	NOx	SO2
Akron, OH	921.87	27	71	59	36	11.4	3243	8.8	42.6	660328	3.34	29560	21	15	59
Albany, NY	997.87	23	72	57	35	11	4281	3.5	50.7	835880	3.14	31458	8	10	39
Allentown, PA	962.35	29	74	54	44	9.8	4260	0.8	39.4	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

http://lib.stat.cmu.edu/DASL/Stories/AirPollutionandMortality.html

Applied Statistical Regression

HS 2011 – Week 04

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.

Applied Statistical Regression
HS 2011 – Week 04

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.*

Matrix Notation

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

$$Y = X\beta + E$$

The elements in this equation are as follows:

→ **see blackboard…**

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*).

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

Hat Matrix Notation

The fitted values are:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

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:

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

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

Applied Statistical Regression
HS 2011 – Week 04

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:

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

What to do if the errors are non-Gaussian?

Applied Statistical Regression
HS 2011 – Week 04

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:

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

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

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 2011 – Week 04

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:

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

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:

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

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 would be to include all predictor variables that are present in the mortality dataset.