Applied Statistical Regression
AS 2014 – Multiple Regression

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 20, 2014

Applied Statistical Regression
AS 2014 – Multiple Regression

What is Regression?

The answer to an everyday question:
How does a target variable of special interest depend on several other (explanatory) factors or causes.

Examples:

- growth of plants, depends on fertilizer, soil quality, …
- apartment rents, depends on size, location, furnishment, …
- car insurance premium, depends on age, sex, nationality, …

Regression:

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

Applied Statistical Regression
AS 2014 – Multiple Regression

Example: Mortality Due to Air Pollution

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	WhiteCllr	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	11.0	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
Boston, MA	934.70	30	74	56	43	12.1	4679	3.5	49.2	2805911	3.23	36644	21	32	62
Bridgeport, CT	899.53	30	73	56	45	10.6	2140	5.3	40.4	438557	3.29	47258	6	4	4
Buffalo, NY	1001.90	24	70	61	36	10.5	6582	8.1	42.5	1015472	3.31	31248	18	12	37
Canton, OH	912.35	27	72	59	36	10.7	4213	6.7	41.0	404421	3.36	29089	12	7	20
Chattanooga, TN	1017.61	42	79	56	52	9.6	2302	22.2	41.3	426540	3.39	25782	18	8	27
Chicago, IL	1024.89	26	76	58	33	10.9	6122	16.3	44.9	606387	3.20	36593	88	63	278
Cincinnati, OH	970.47	34	77	57	40	10.2	4101	13.0	45.7	1401491	3.21	31427	26	26	146
Cleveland, OH	985.95	28	71	60	35	11.1	3042	14.7	44.6	1898825	3.29	35720	31	21	64
Columbus, OH	958.84	31	75	58	37	11.9	4259	13.1	49.6	124833	3.26	29761	23	9	15
Dallas, TX	860.10	46	85	54	35	11.8	1441	14.8	51.2	1957378	3.22	38769	1	1	1

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

Multiple Linear Regression

We use linear modeling for a multiple predictor regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + E_i$$

- there are now p predictors
- the problem cannot be visualized in a scatterplot
- there will be n observations of response and predictors
- goal: estimating the coefficients $\beta_0, \beta_1, ..., \beta_p$ from the data

IMPORTANT: simple linear regression of the response on each of the predictors does not equal multiple regression, where *all predictors are used simultanously*.

Applied Statistical Regression
AS 2014 – Multiple Regression

Data Preparation: Visualization

Because we cannot inspect the data in a xy-scatterplot, data visualization and data preparation becomes an important task. We need to identify the necessary variable transformations, mitigate the effect of outliers, …

Step 1: Plotting the marginal distribution (i.e. histograms)

```
> par(mfrow=c(4,4))
> for (i in 1:15) hist(apm[,i], main= "...")
```

Step 2: Identify erroneous and missing values

```
> any(is.na(apm))
[1] FALSE
```

Data Preparation: Transformations

Linear regression and its output are easier to comprehend if one is using an intuitive scale for the variables. Please note that linear transformations do not change the results. However, any non-linear transformation will do so.

Step 3: linear transformations $x' = ax + b$

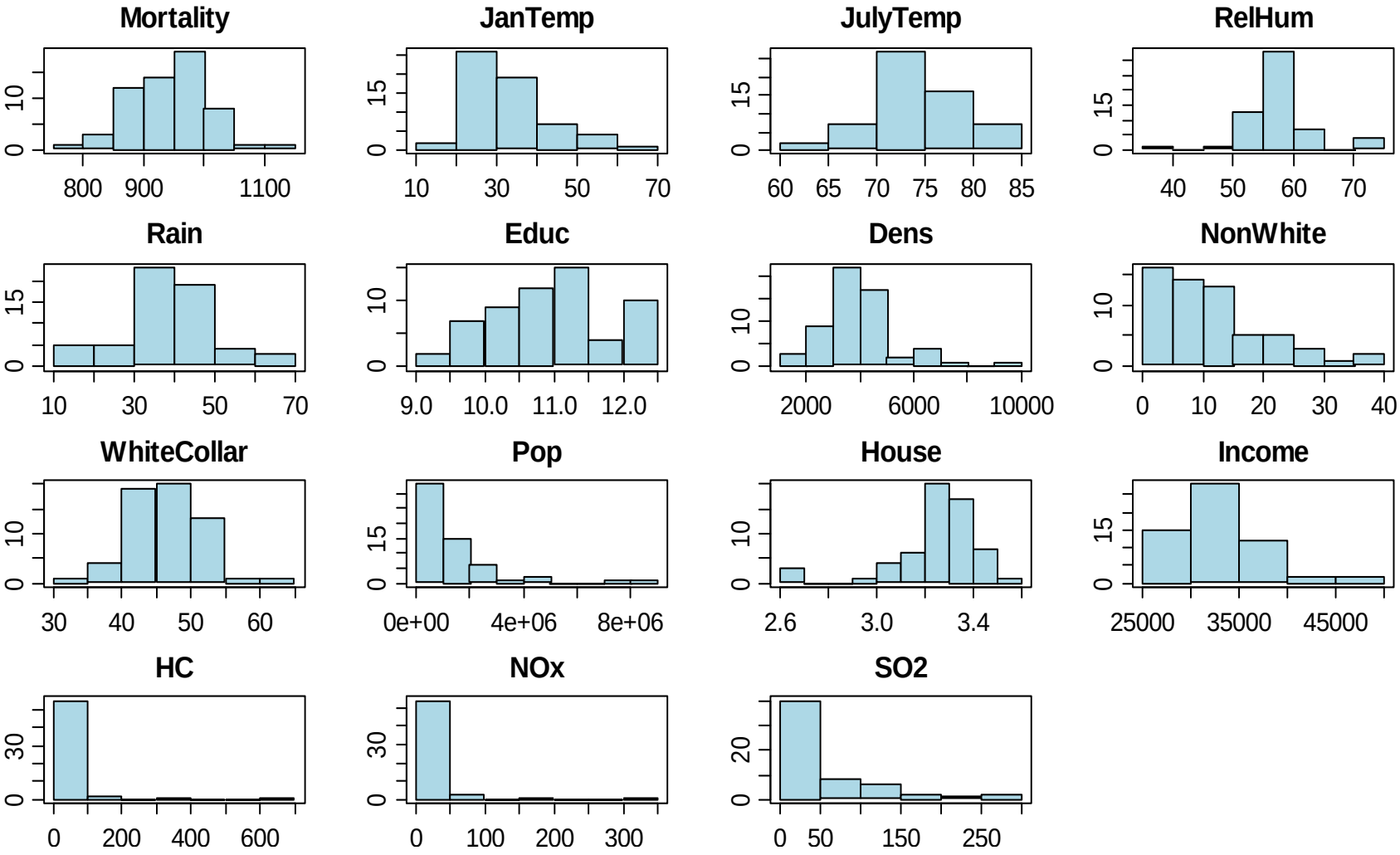
```
> apm$JanTemp  <- (5/9)*(apm$JanTemp-32)
> apm$JulyTemp <- (5/9)*(apm$JulyTemp-32)
> apm$Rain     <- (2.54)*apm$Rain
```

Step 4: log-transformation $x' = \log(x)$

For all variables where it is necessary/beneficial...

Applied Statistical Regression
AS 2014 – Multiple Regression

Data Preparation: Transformations



Why Simple Regression Is Not Enough

Performing many simple lineare regressions of the response on any of the predictors is not the same as multiple regression!

Observation	x1	x2	yy
1	0	-1	1
2	1	0	2
3	2	1	3
4	3	2	4
5	0	1	-1
6	1	2	0
7	2	3	1
8	3	4	2

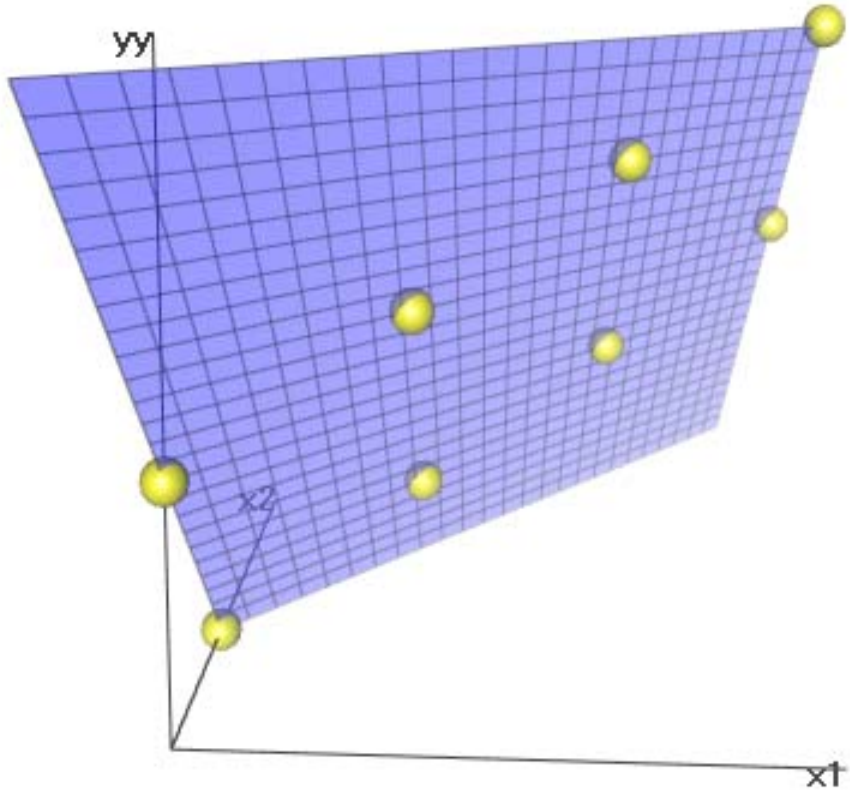
We have $y_i = \hat{y}_i = 2x_{i1} - x_{i2}$, i.e. a perfect fit.
Hence, all residuals are zero and we estimate $\hat{\sigma}_E^2 = 0$.

→ *The result can be visualized with a 3d-plot!*

Applied Statistical Regression
AS 2014 – Multiple Regression

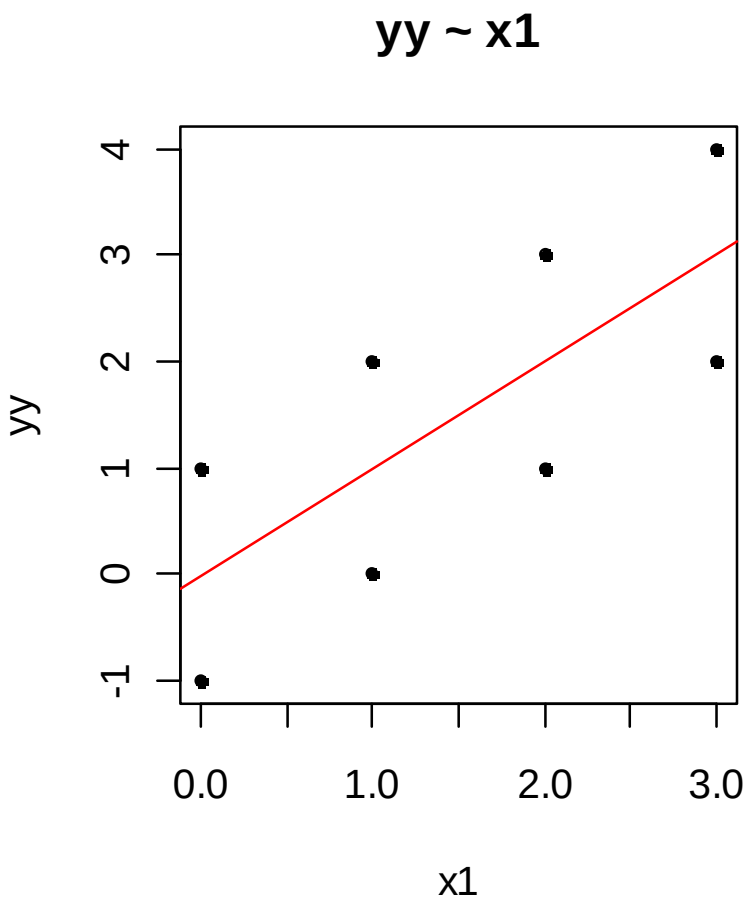
Why Simple Regression Is Not Enough

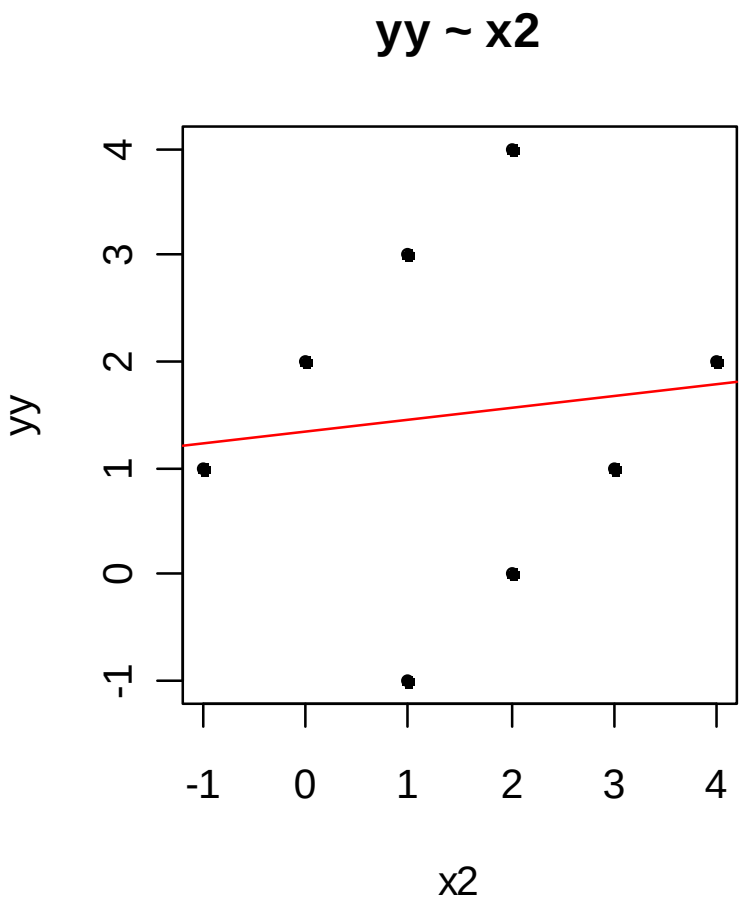
```
> library(Rcmdr)
> scatter3d(yy ~ x1 + x2, axis.scales=FALSE)
```



Applied Statistical Regression
AS 2014 – Multiple Regression

Why Simple Regression Is Not Enough





The Multiple Linear Regression Model

In colloquial notation, the model is:

$$Mortality_i = \beta_0 + \beta_1 \cdot JanTemp_i + \beta_2 \cdot JulyTemp_i + ... + \beta_{14} \cdot \log(SO_2)_i + E_i$$

More generally, the multiple linear regression model specifies the relation between response y and predictors $x_1, ..., x_p$. There are observations $i = 1, ..., n$. We use the double index notation:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + E_i \quad \text{for} \quad i = 1, ..., n$$

Here, β_0 is the intercept and $\beta_1, ..., \beta_p$ are regression coefficients.

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

Matrix Notation

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

$$y = X\beta + E$$

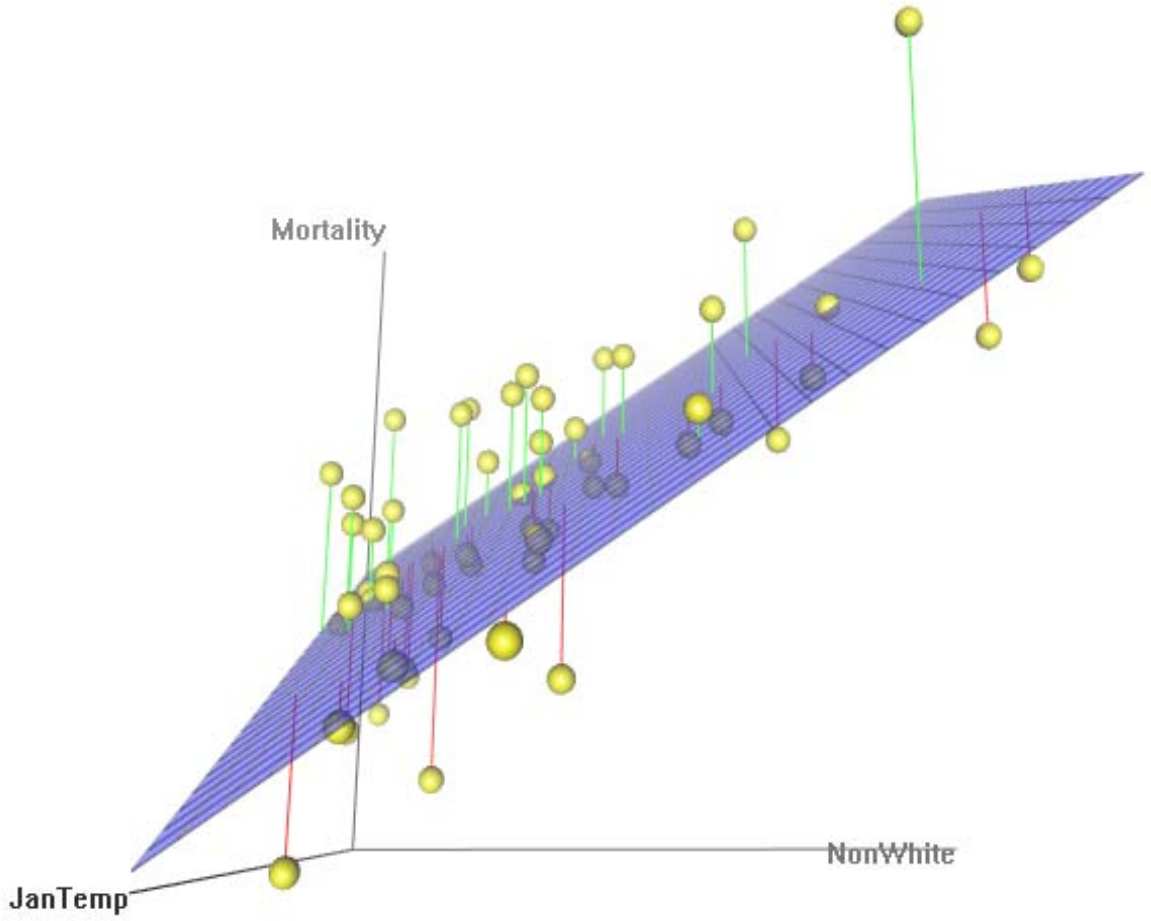
The elements in this equation are as follows:

→ **see blackboard…**

Applied Statistical Regression
AS 2014 – Multiple Regression

Fitting Multiple Regression Models

Toy example: $Mortality_i = \beta_0 + \beta_1 \cdot JanTemp_i + \beta_2 \cdot NonWhite_i + E_i$



Least Squares Algorithm

The *paradigm* is to determine the regression coefficients such that the *sum of squared residuals is minimal*. This amounts to minimizing the quality function:

$$Q(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))^2$$

We can take partial derivatives with respect to $\beta_0, \beta_1, ..., \beta_p$ and so obtain a linear equation system with $(p+1)$ unknowns and the same number of equations.

→ **Mostly (but not always...), there is a unique solution.**

Normal Equations and Their Solutions

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

$$(X^T X)\beta = X^T y \quad \text{resp.} \quad \hat{\beta} = (X^T X)^{-1} X^T y = Hy$$

- 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!

Multiple Regression in R

In R, multiple linear least squares regression is carried out with command `lm()`. The syntax is as follows:

```
fit <- lm(Mortality ~ JanTemp + JulyTemp + RelHum +
                      Rain + Educ + Dens + NonWhite +
                      WhiteCollar + log(Pop) + House +
                      Income + log(HC) + log(NOx) +
                      log(SO2), data=apm)
```

An often useful short notation is:

```
fit <- lm(Mortality ~ ., data=apm)
```

Except for the response, all variables in `apm` are predictors.

Estimating the Error Variance

For producing confidence intervals for the coefficients, testing the regression coefficients and producing a prediction interval for future observation, having an estimate of the error variance is indispensable.

$$\hat{\sigma}_E^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} r_i^2$$

The estimate is given by the “average residual”. The division by $n-(p+1)$ is for obtaining an unbiased estimator. Here, p ïs the number of predictors, and $(p+1)$ is the number of parameters which are estimated.

Assumptions on the Error Term

The 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
- $E_i \sim N(0, \sigma_E^2)$, the errors are normally distributed

Note: As in simple linear regression, we do not require Gaussian distribution for OLS estimation and certain optimality results, i.e. the Gauss-Markov theorem.

But: All tests and confidence intervals rely on the Gaussian, and there are better estimates for non-normal data

Applied Statistical Regression
AS 2014 – Multiple 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!)

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?

Benefits of Linear Regression

- **Inference on the relation between y and $x_1, ..., x_p$**

The goal is to understand if and how strongly the response variable depends on the predictor. There are performance indicators as well as statistical tests adressing the issue.

- **Prediction of (future) observations**

The regression equation can be employed to predict the response value for any given predictor configuration.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

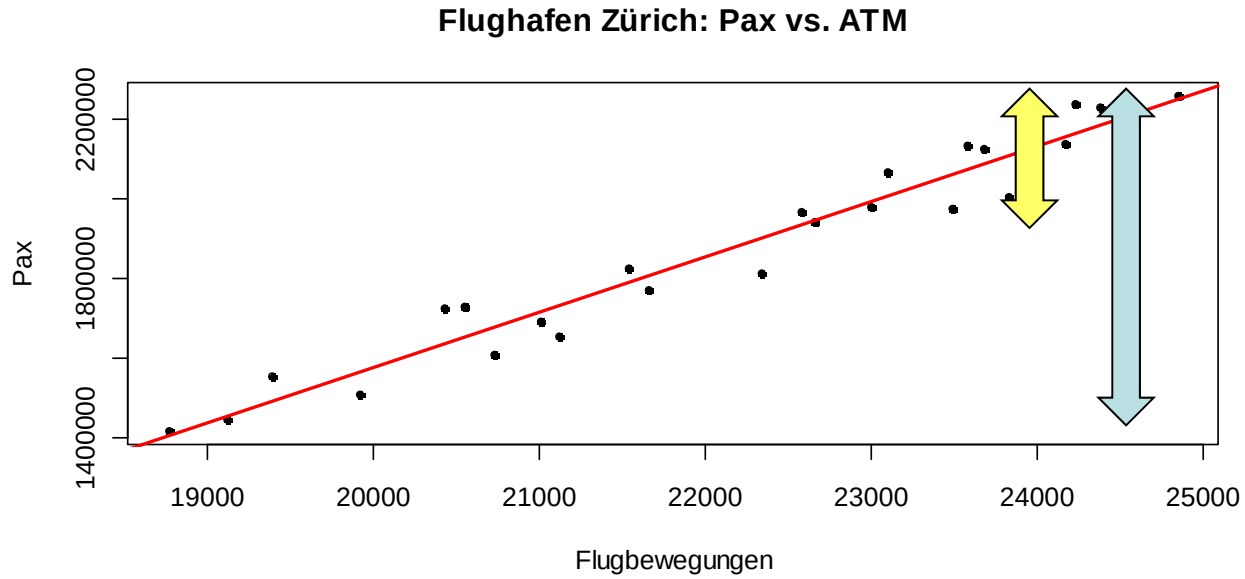
However, this mostly will not work well for extrapolation!

Applied Statistical Regression
AS 2014 – Multiple Regression

R^2: *The Coefficient of Determination*

The coefficient of determination R^2 tells which portion of the total variation is accounted for by the regression hyperplane.

→ For multiple linear regression, visualization is impossible!
→ The number of predictor used should be taken into account.



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:

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

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!**

The adjusted coefficient of determination is defined as:

$$adjR^2 = 1 - \frac{n-1}{n-(p+1)} \cdot (1-R^2) \in [0,1]$$

Hence, the adjusted R-squared is always (but in many cases irrelevantly) smaller than the plain R-squared. The biggest discrepancy is with small n, large p and small R^2.

Applied Statistical Regression
AS 2014 – Multiple Regression

Confidence Interval for Coefficient β_j

We can give a 95%-CI for the regression coefficient β_j.
It tells which values, besides the point estimate $\hat{\beta}_j$, are plausible too.

Note: This uncertainty comes from sampling effects

95%-VI for β_j : $\hat{\beta}_j \pm qt_{0.975;n-(p+1)} \cdot \hat{\sigma}_{\hat{\beta}_j}$

In R:

```
> fit <- lm(Mortality ~ ., data=mt)
> confint(fit, "Educ")
       2.5 %   97.5 %
Educ -31.03177 4.261925
```

Testing the Coefficient β_j

There is a statistical hypothesis test which can be used to check whether $\hat{\beta}_j$ is significantly different from zero, or different from any other arbitrary value b. The null hypothesis is:

$$H_0 : \beta_j = 0, \text{ resp. } H_0 : \beta_j = b$$

One usually tests two-sided on the 95%-level. The alternative is:

$$H_A : \beta_j \neq 0, \text{ resp. } H_A : \beta_j \neq b$$

As a test statistic, we use:

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}, \text{ resp. } T = \frac{\hat{\beta}_j - b}{\hat{\sigma}_{\hat{\beta}_j}}, \text{ both follow a } t_{n-(p+1)} \text{ distribution.}$$

Applied Statistical Regression
AS 2014 – Multiple Regression

Reading R-Output

```
> summary(fit.orig)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1496.4915   572.7205   2.613  0.01224 *
JanTemp       -2.4479     0.8808  -2.779  0.00798 **
...
Dens          11.9490    16.1836   0.738  0.46423
NonWhite     326.6757    62.9092   5.193 5.09e-06 ***
WhiteCollar -146.3477   112.5510  -1.300  0.20028
...
---
Residual standard error: 34.23 on 44 degrees of freedom
Multiple R-squared: 0.7719, Adjusted R-squared: 0.6994
F-statistic: 10.64 on 14 and 44 DF,  p-value: 6.508e-10
```

Note: due to space constraints, this is only a part of the output!

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 I 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!*

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:

c) The p-values of the individual hypothesis tests are based on the assumption that the other predictors remain in the model and do not change. Therefore, you must not drop more than one single non-significant predictor at a time!

Solution: *drop one*, *re-evaluate the model*, *drop one*, ...

Applied Statistical Regression
AS 2014 – Multiple Regression

Simple Variable Selection

Goal: Dropping all predictors from the regression model which are not necessary, i.e. do not show a significant impact on the response.

How: In a step-by-step manner, the least significant predictor is dropped from the model, as long as its p-value still exceeds the value of 0.05.

In R:
```
> fit <- update(fit, . ~ . - RelHum)
> summary(fit)
```

→ **Exercise: try do to this for the Mortality Data**

Comparing Hierachical Models

Idea: Correctly comparing two multiple linear regression models when the smaller has >1 predictor less than the bigger.

Where and why do we need this?

- for the 3 pollution variables in the mortality data.
- soon also for the so-called factor/dummy variables.

Idea: We compare the residual sum of squares (RSS):

Big model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q + \beta_{q+1} x_{q+1} + ... + \beta_p x_p$

Small model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$

The big model must contain all the predictors from the small model, else they are not hierarchical and the test does not apply.

Comparing Hierarchical Models

Null hypothesis:

$H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_p = 0$, versus the alternative hypothesis that at least one $\beta_j \neq 0,\ j = q+1,...p$

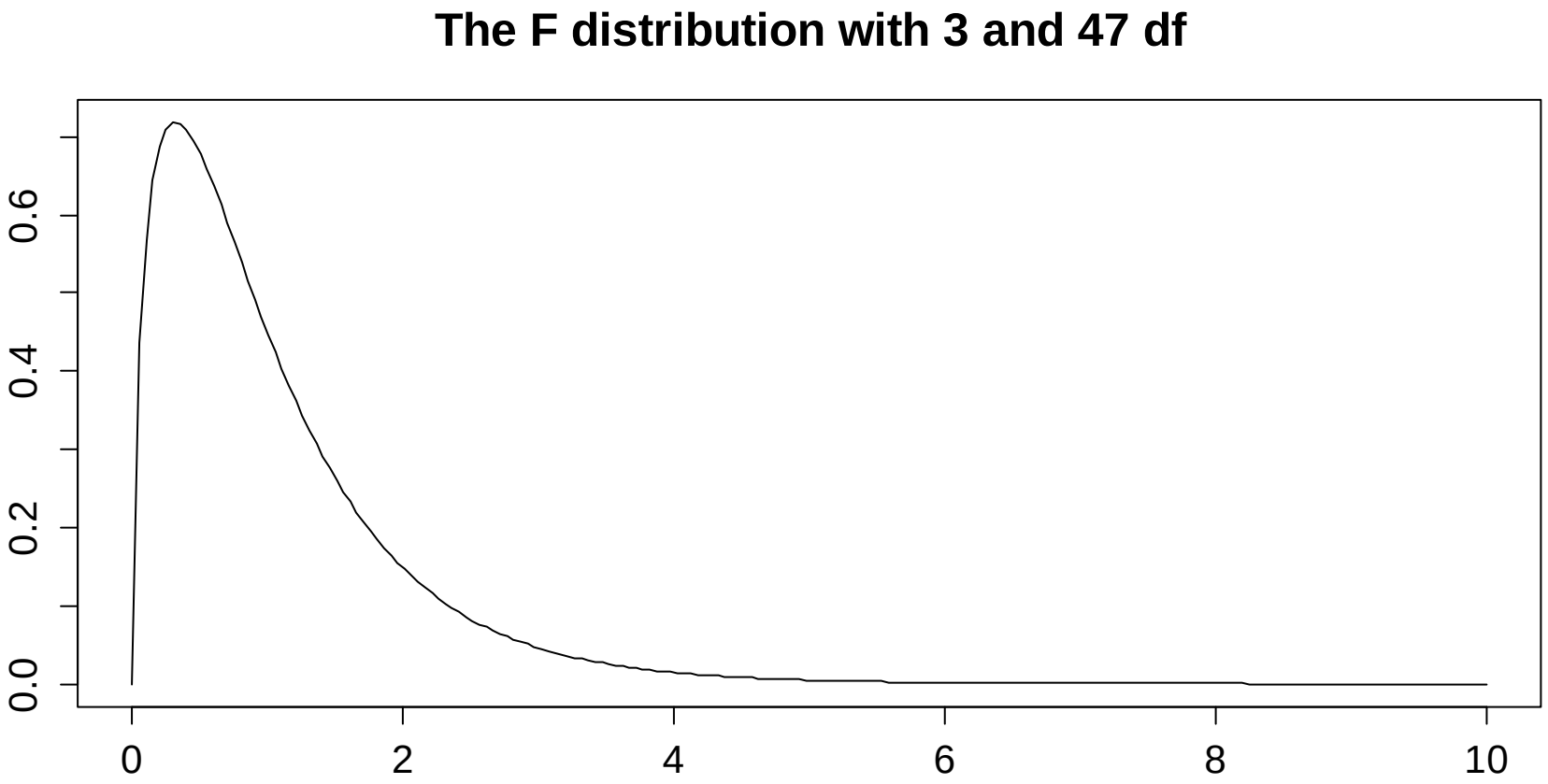
The test compares the RSS of the big and the small model:

$$F = \frac{n-(p+1)}{p-q} \cdot \frac{RSS_{Small} - RSS_{Big}}{RSS_{Big}} \sim F_{p-q,n-(p+1)}$$

→ If the F-value is small ($p \geq 0.05$), the two models perform equally. There is no evidence against the null and we can continue working with the small model.

Applied Statistical Regression
AS 2014 – Multiple Regression

Density Function of the F-distribution



Applied Statistical Regression
AS 2014 – Multiple Regression

Comparing Hierachical Models in R

```
> f.big   <- lm(Mortality ~ ., data=mt)
> f.small <- update(f.big, .~.-log(HC)-log(Nox)-log(SO2))

> anova(f.big, f.small)

Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain +
         Educ + Dens + NonWhite + WhiteCollar + log(Pop) +
         House + Income + log(HC) + log(Nox) + log(SO2)
Model 2: Mortality ~ JanTemp + JulyTemp + RelHum + Rain +
         Educ + Dens + NonWhite + WhiteCollar + log(Pop) +
         House + Income

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     44 51543
2     47 61244 -3   -9700.8 2.7604 0.0533 .
```

The Global F-Test

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

This is another hierachical model comparison. The full model is tested against a small model with only the intercept, but without any predictors.

We are testing the null $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ against the alternative $H_A : \beta_j \neq 0$ for at least one predictor x_j. This test is again based on comparing the RSS:

$$F = \frac{n-(p+1)}{p} \cdot \frac{RSS_{Small} - RSS_{Big}}{RSS_{Big}} \sim F_{p,n-(p+1)}$$

→ **Test statistic and p-value are shown in the R summary!**

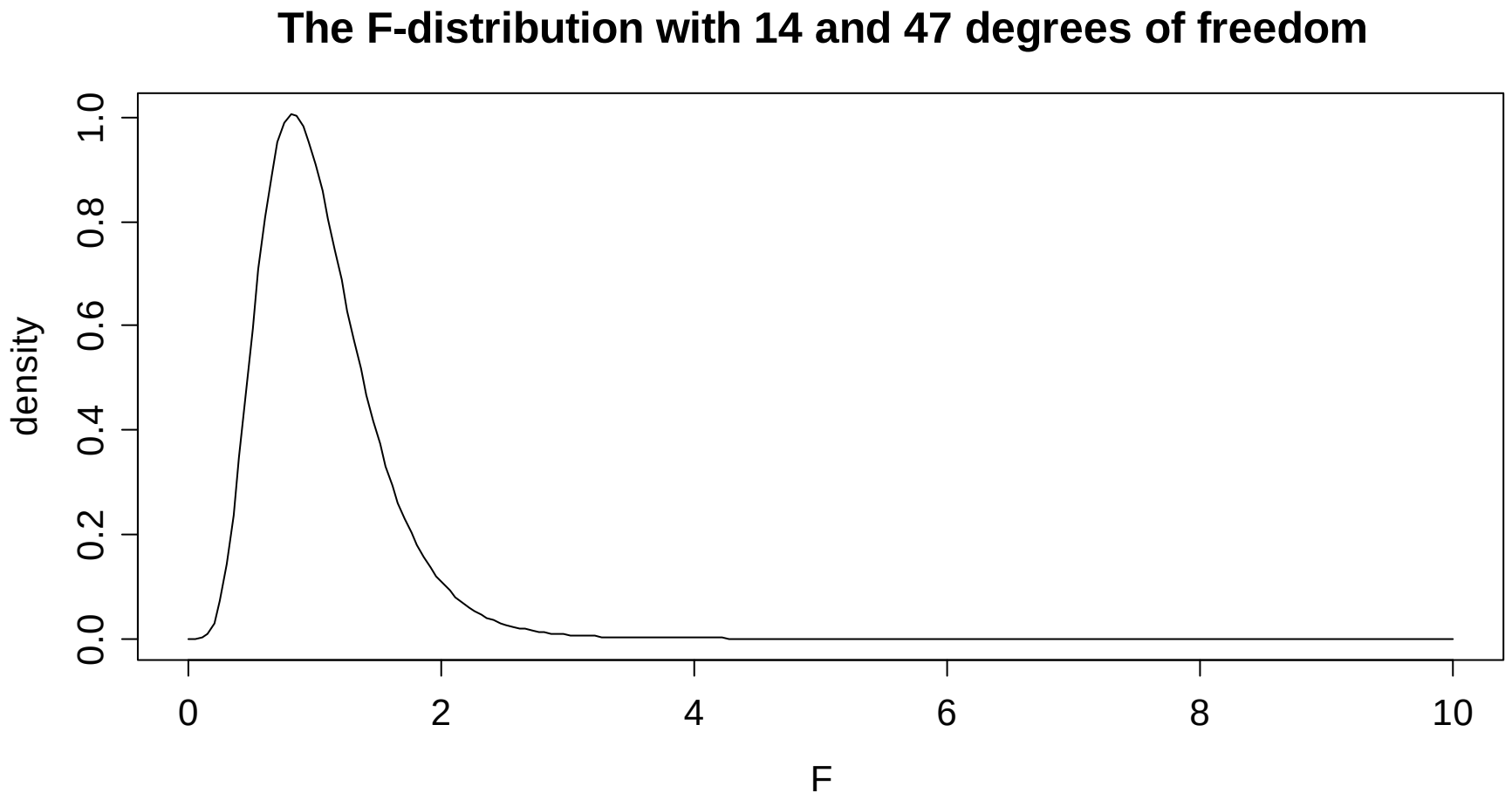
Applied Statistical Regression
AS 2014 – Multiple Regression

Reading R-Output

```
> summary(fit.orig)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1496.4915   572.7205   2.613  0.01224 *
JanTemp       -2.4479     0.8808  -2.779  0.00798 **
...
Dens          11.9490    16.1836   0.738  0.46423
NonWhite     326.6757    62.9092   5.193 5.09e-06 ***
WhiteCollar -146.3477   112.5510  -1.300  0.20028
...
---
Residual standard error: 34.23 on 44 degrees of freedom
Multiple R-squared: 0.7719, Adjusted R-squared: 0.6994
F-statistic: 10.64 on 14 and 44 DF,  p-value: 6.508e-10
```

Note: due to space constraints, this is only a part of the output!

Density Function of the F-distribution



Prediction

The regression equation can be employed for predicting the response value in any given predictor configuration.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{.1} + \hat{\beta}_2 x_{.2} + \ldots + \hat{\beta}_p x_{.p}$$

Note:
This can be a predictor configuration that was not part of the original data. For example a (new) city, for which only the predictors are known, but the mortality is not.

Be careful:
Only interpolation, i.e. prediction within the range of observed y-values works well, extrapolation yields non-reliable results.

Prediction in R

We can use the regression fit for predicting new observations. The syntax is as follows

```
> fit.big <- lm(Mortality ~ ., data=mt)
> dat       <- data.frame(JanTemp=..., ...)
> predict(fit.big, newdata=dat)
1 932.488
```

The x-values need to be provided in a data frame. The variable (column) names need to be identical to the predictor names. Of course, all predictors need to be present.

Then, it is simply applying the `predict()`-procedure.

Confidence- and Prediction Interval

The confidence interval for the fitted value and the prediction interval for future observation also exist in multiple regression.

a) 95%-CI for the fitted value $E[y \mid x]$

```
> predict(fit, newdata=dat, "confidence")
```

b) 95%-PI for a future observation $\hat{y}$:

```
> predict(fit, newdata=dat, "prediction")
```

- The visualization of these intervals is no longer possible in the case of multiple regression
- It is possible to write explicit formulae for the intervals using the matrix notation. We omit them here.

Versatility of Multiple Linear Regression

Despite that we are using linear models only, we have a versatile and powerful tool. While the response is always a continuous variable, different predictor types are allowed:

- **Continuous Predictors**
  Default case, e.g. *temperature*, *distance*, *pH-value*, …
- **Transformed Predictors**
  For example: $log(x)$, $sqrt(x)$, $arcsin(\sqrt{x})$,...
- **Powers**
  We can also use: x^{-1}, x^2, x^3, ...
- **Categorical Predictors**
  Often used: *sex*, *day of week*, *political party*, …

Categorical Predictors

The canonical case in linear regression are *continuous predictor variables* such as for example:

→ *temperature, distance, pressure, velocity, ...*

While in linear regression, we *cannot have categorical response*, it is perfectly valid to have *categorical predictors*:

→ *yes/no, sex (m/f), type (a/b/c), shift (day/evening/night), ...*

Such categorical predictors are often also called **factor variables**. In a linear regression, each level of such a variable is encoded by a dummy variable, so that $(\ell - 1)$ degrees of freedom are spent.

Regression with a Factor Variable

The lathe (*in German: Drehbank*) **dataset:**

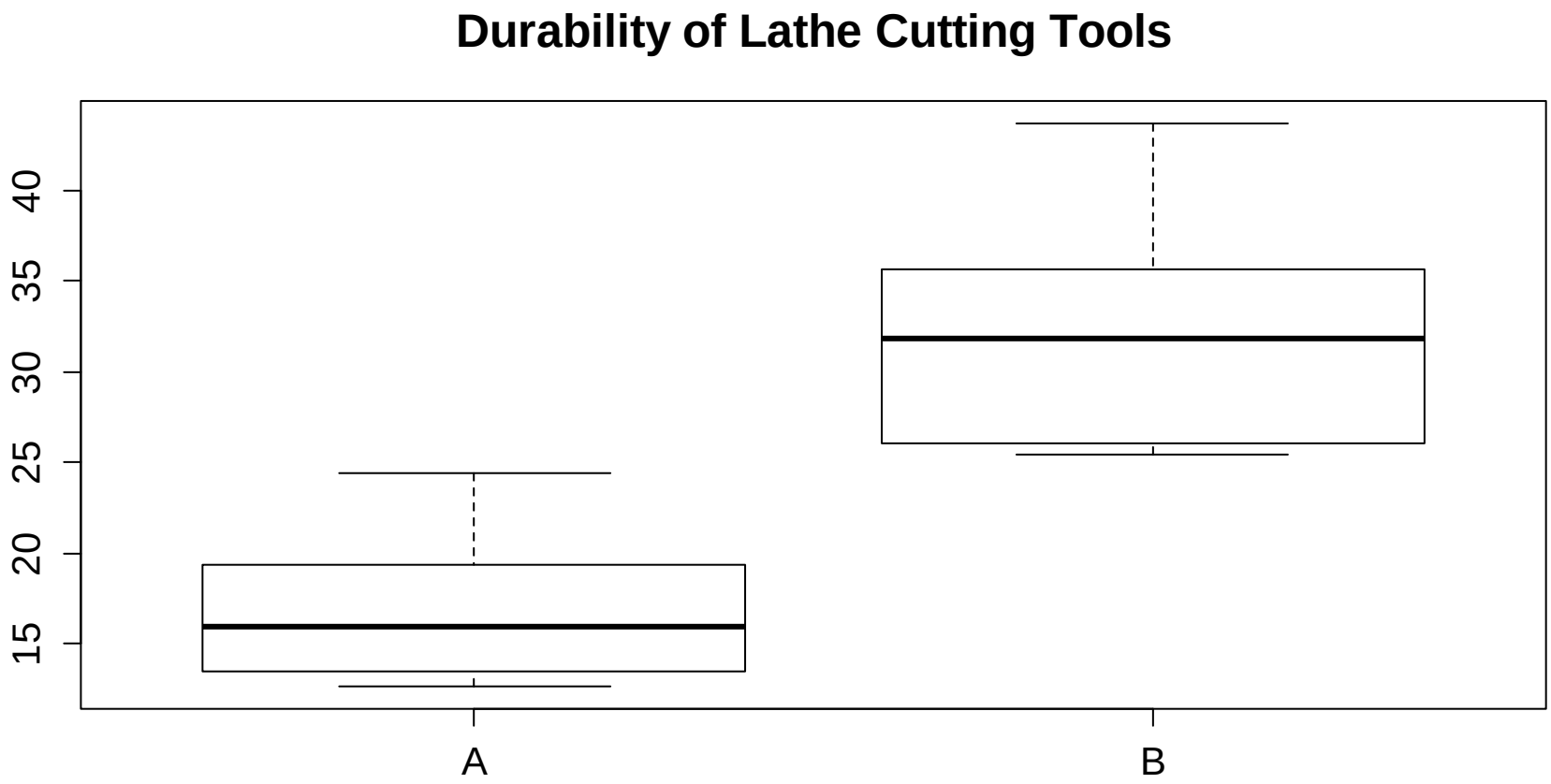
- y_i life time of cutting tool i
- $\tilde{x}_i$ type of tool i, A or B

Dummy variable encoding:

$$x_i = \begin{cases} 0 & tool\ \ type\ \ A \\ 1 & tool\ \ type\ \ B \end{cases}$$

Applied Statistical Regression
AS 2014 – Multiple Regression

Typical Visualization of a Factor Model



Applied Statistical Regression
AS 2014 – Multiple Regression

Interpretation of the Factor Model

→ **See blackboard…**

```
> summary(fit)

Call: lm(formula = hours ~ tool, data = lathe)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.110      1.628  10.508 4.14e-09 ***
toolB         14.818      2.303   6.435 4.68e-06 ***
---
Residual standard error: 5.149 on 18 degrees of freedom
Multiple R-squared: 0.697,  Adjusted R-squared: 0.6802
F-statistic: 41.41 on 1 and 18 DF,  p-value: 4.681e-06
```

Applied Statistical Regression
AS 2014 – Multiple Regression

Another View: t-Test

→ **The 1-factor-model is a t-test for non-paired data!**

```
> t.test(hours ~ tool, data=lathe, var.equal=TRUE)

Two Sample t-test

data:  hours by tool
t = -6.435, df = 18, p-value = 4.681e-06
alternative hypothesis: true diff in means is not 0
95 percent confidence interval:
 -19.655814  -9.980186
sample estimates:
mean in group A mean in group B
         17.110          31.928
```

Example: Binary Categorical Variable

The lathe (*in German: Drehbank*) **dataset:**

- y lifetime of a cutting tool in a turning machine
- x_1 speed of the machine in rpm
- $\tilde{x}_2$ tool type A or B

Dummy variable encoding:

$$x_2 = \begin{cases} 0 & tool\ \ type\ \ A \\ 1 & tool\ \ type\ \ B \end{cases}$$

Applied Statistical Regression
AS 2014 – Multiple Regression

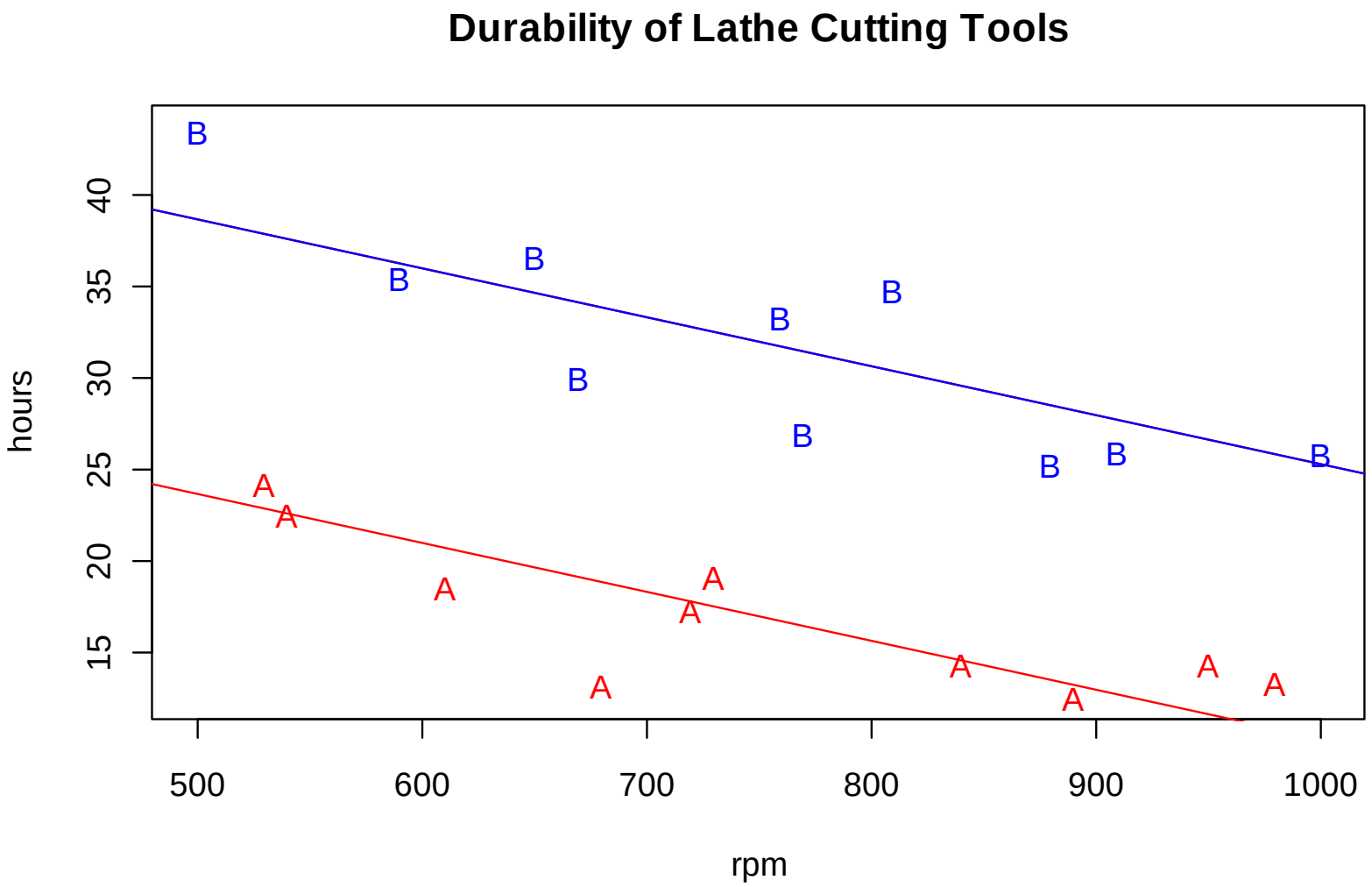
Interpretation of the Model

→ **see blackboard…**

```
> summary(lm(hours ~ rpm + tool, data = lathe))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
rpm         -0.02661    0.00452  -5.887 1.79e-05 ***
toolB       15.00425    1.35967  11.035 3.59e-09 ***
---
Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared: 0.9003,  Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,   p-value: 3.086e-09
```

Applied Statistical Regression
AS 2014 – Multiple Regression

The Dummy Variable Fit



A Model with Interactions

Question: do the slopes need to be identical?

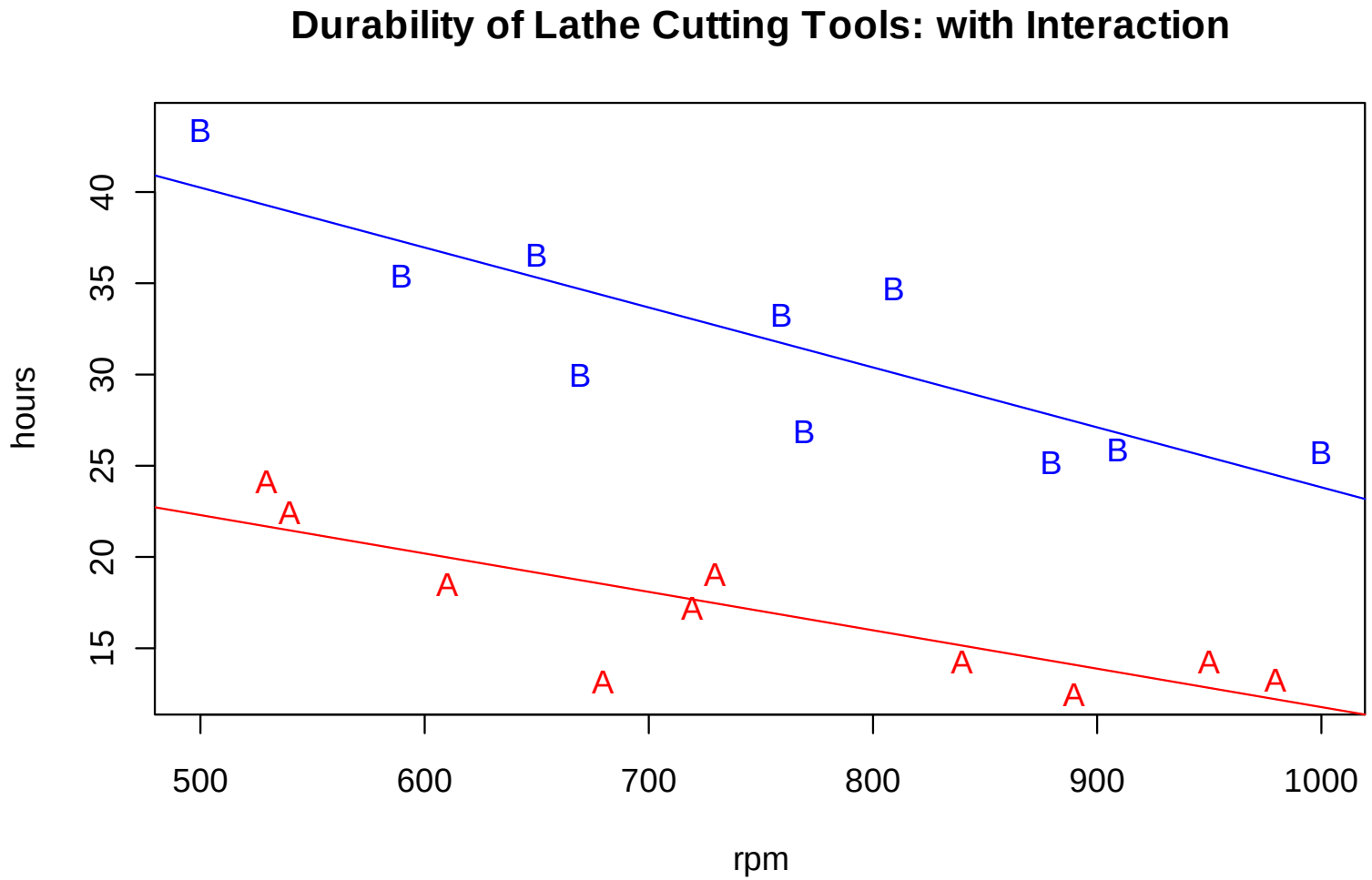
→ with the appropriate model, the answer is no!

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + E$$

→ **see blackboard for model interpretation…**

Applied Statistical Regression
AS 2014 – Multiple Regression

Different Slopes for the Regression Lines



Applied Statistical Regression
AS 2014 – Multiple Regression

Summary Output

```
> summary(lm(hours ~ rpm + tool + rpm:tool, data=lathe))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
rpm         -0.020970   0.006074  -3.452  0.00328 **
toolB       23.970593   6.768973   3.541  0.00272 **
rpm:toolB   -0.011944   0.008842  -1.351  0.19553
---
Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared: 0.9105,  Adjusted R-squared: 0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

How Complex the Model Needs to Be?

Question 1: do we need different slopes for the two lines?

$$H_0 : \beta_3 = 0 \quad \text{against} \quad H_A : \beta_3 \neq 0$$

→ no, see individual test for the interaction term on previous slide!

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{against} \quad H_A : \beta_2 \neq 0 \;\; and / or \;\; \beta_3 \neq 0$$

→ this is a hierarchical model comparison
→ we try to exclude interaction and dummy variable together

R offers convenient functionality for this test, see next slide!

Testing the Tool Type Variable

Hierarchical model comparison with anova( ):

```
> fit.small <- lm(hours ~ rpm, data=lathe)
> fit.big <- lm(hours ~ rpm + tool + rpm:tool, data=lathe)
> anova(fit.small, fit.big)
Model 1: hours ~ rpm
Model 2: hours ~ rpm + tool + rpm:tool
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 1282.08
2     16  140.98  2    1141.1 64.755 2.137e-08 ***
```

→ The bigger model, i.e. making a distinction between the tools, is significantly better. The main effect is enough, though.

Categorical Input with More Than 2 Levels

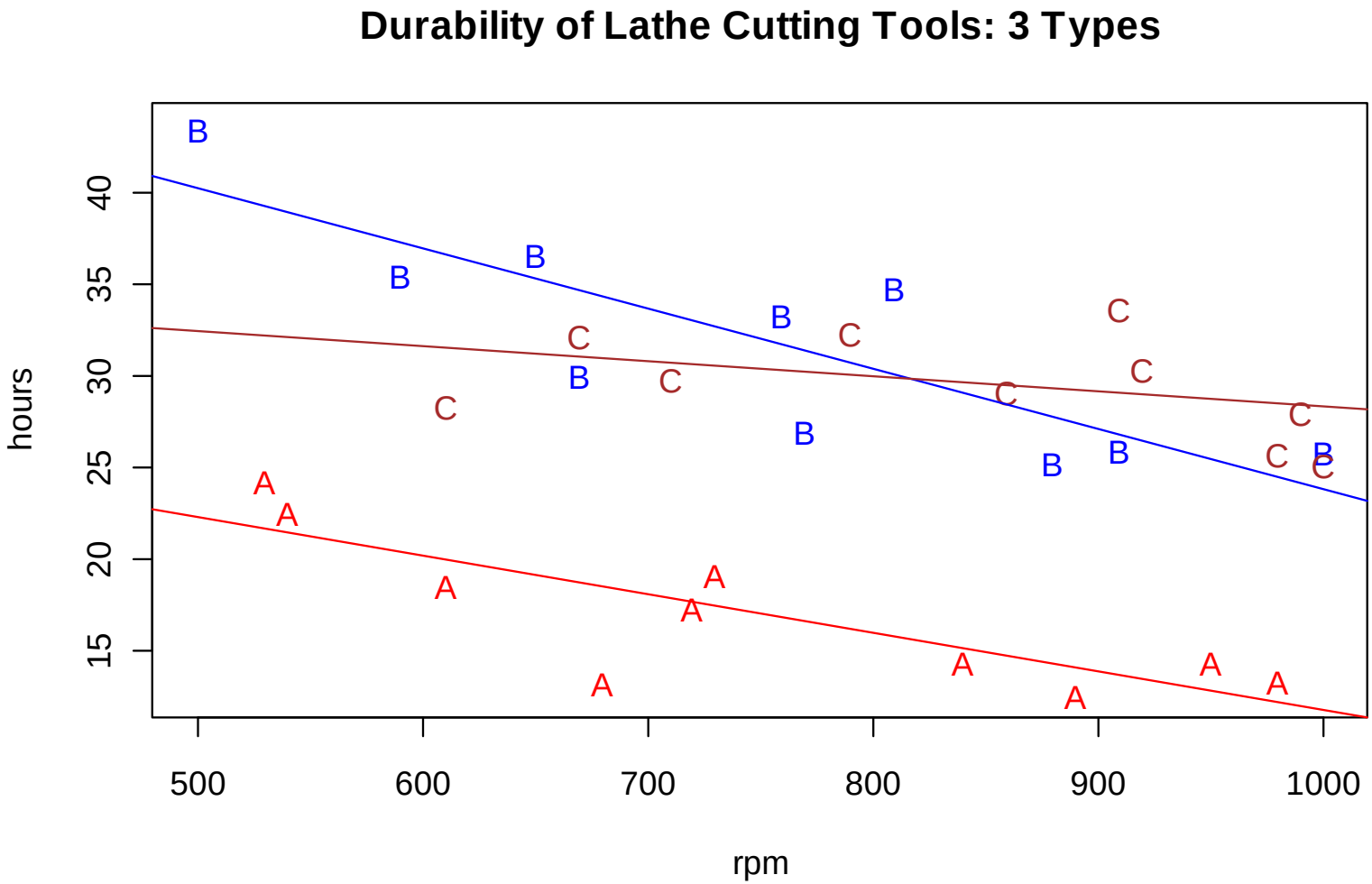
There are now 3 tool types A, B, C:

x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

Main effect model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + E$

With interactions: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + E$

Three Types of Cutting Tools



Applied Statistical Regression
AS 2014 – Multiple Regression
Summary Output

```
> summary(lm(hours ~ rpm + tool + rpm:tool, data = abc.lathe)

Coefficients:Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.496024   7.290 1.57e-07 ***
rpm         -0.020970   0.005894  -3.558  0.00160 **
toolB       23.970593   6.568177   3.650  0.00127 **
toolC        3.803941   7.334477   0.519  0.60876
rpm:toolB   -0.011944   0.008579  -1.392  0.17664
rpm:toolC    0.012751   0.008984   1.419  0.16869
---
Residual standard error: 2.88 on 24 degrees of freedom
Multiple R-squared: 0.8906,    Adjusted R-squared: 0.8678
F-statistic: 39.08 on 5 and 24 DF,  p-value: 9.064e-11
```

This summary is of limited use for deciding about model complexity. We require hierarchical model comparisons!

Applied Statistical Regression
AS 2014 – Multiple Regression

Inference with Categorical Predictors

Do not perform individual hypothesis tests on factors that have more than 2 levels, they are meaningless!

Question 1: do we have different slopes?

$$H_0 : \beta_4 = 0 \ \ and \ \ \beta_5 = 0 \text{ against } H_A : \beta_4 \neq 0 \ \ and/or \ \ \beta_5 \neq 0$$

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \text{ against } H_A : any \ of \ \ \beta_2, \beta_3, \beta_4, \beta_5 \neq 0$$

→ Again, R provides convenient functionality: `anova( )`

Applied Statistical Regression
AS 2014 – Multiple Regression

Anova Output

```
> anova(fit.abc)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
rpm        1  139.08  139.08 16.7641  0.000415 ***
tool       2 1422.47  711.23 85.7321 1.174e-11 ***
rpm:tool   2   59.69   29.84  3.5974  0.043009 *
Residuals 24  199.10    8.30
```

→ The interaction term is weakly significant. Thus, there is some weak evidence for the necessity of different slopes.

→ The p-value for the tool variable includes omitting interaction and main effect. Being strongly significant, we have strong evidence that tool type distinction is needed.

Applied Statistical Regression
AS 2014 – Multiple Regression

Residual Analysis – Model Diagnostics

Why do it? And what is it good for?

a) To make sure that estimates and inference are valid

- $E[E_i] = 0$
- $Var(E_i) = \sigma_E^2$
- $Cov(E_i, E_j) = 0$
- $E_i \sim N(0, \sigma_E^2 I),\ i.i.d$

b) Identifying unusual observations

Often, there are just a few observations which "are not in accordance" with a model. However, these few can have strong impact on model choice, estimates and fit.

Residual Analysis – Model Diagnostics

Why do it? And what is it good for?

c) Improving the model

- Transformations of predictors and response
- Identifying further predictors or interaction terms
- Applying more general regression models

- There are both model diagnostic graphics, as well as numerical summaries. The latter require little intuition and can be easier to interpret.

- However, the graphical methods are far more powerful and flexible, and are thus to be preferred!

Residuals vs. Errors

All requirements that we made were for the errors E_i. However, they cannot be observed in practice. All that we are left with are the residuals r_i , which are only estimates of the errors.

But:

- The residuals r_i do share some properties of the errors E_i , but not all – there are some important differences.
- In particular, even in cases where the E_i are uncorrelated and have constant variance, the residuals r_i feature some estimation-related correlation and non-constant variance.

→ *Does residual analysis make sense?*

Standardized/Studentized Residuals

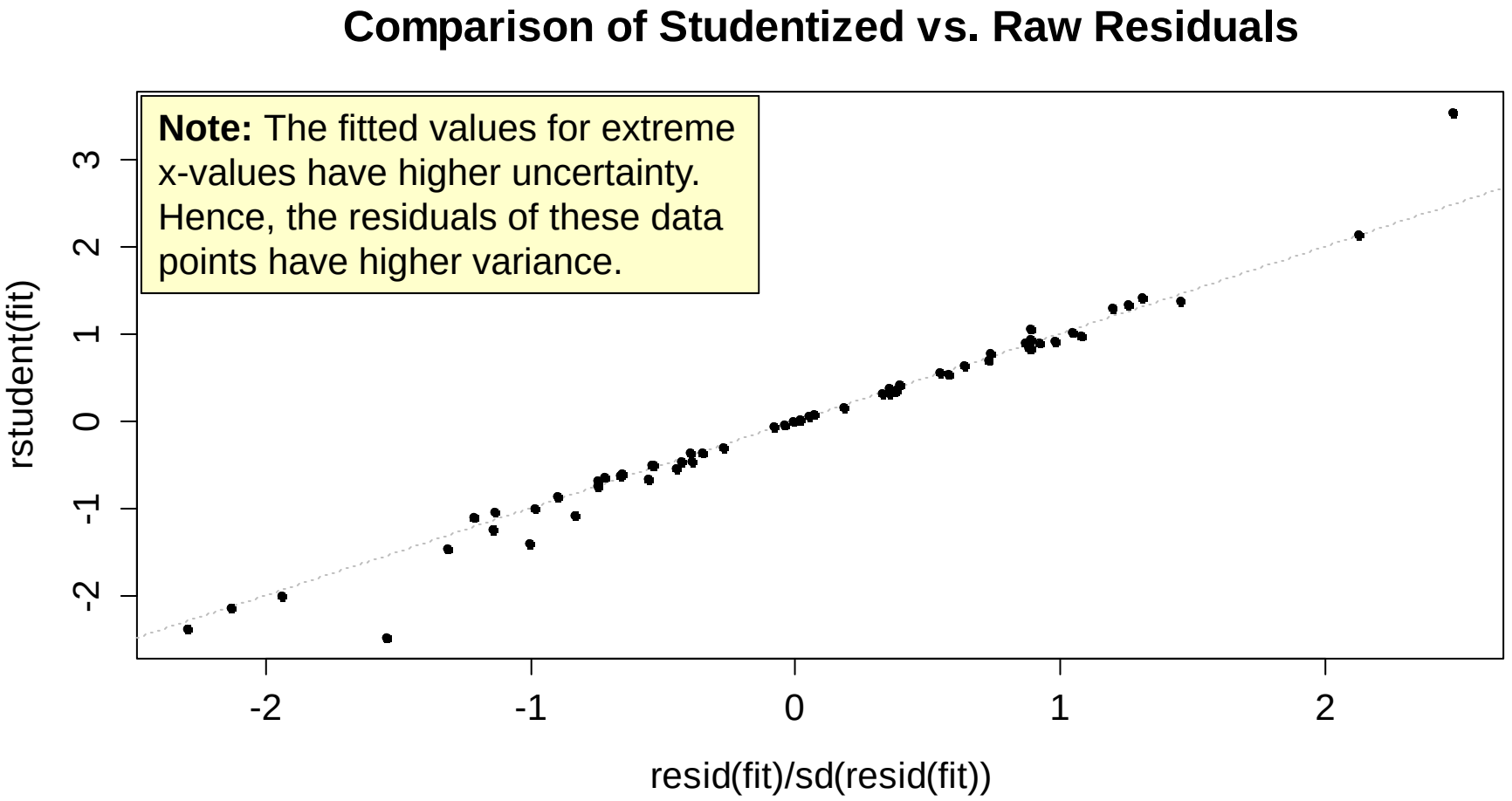
- The *estimation-induced* correlation and non-constant variance in the residuals is usually very small. Thus, residual analysis using the raw residuals r_i is both useful and sensible.
- One can try to improve the raw residual r_i with dividing it by an estimate of its standard deviation.

$$\tilde{r}_i = \frac{r_i}{\hat{\sigma}_E \cdot \sqrt{1 - h_{ii}}}, \quad h_{ii} \text{ is the diagonal element of hat matrix}$$

If $\hat{\sigma}_E$ is the residual standard error, we speak of *standardized residuals*. Sometimes, one also uses a different estimate $\hat{\sigma}_E$ that was obtained by ignoring the i^{th} datapoint. One then speaks of *studentized residuals*.

Applied Statistical Regression
AS 2014 – Multiple Regression

Studentized vs. Raw Residuals



Toolbox for Model Diagnostics

There are 4 "standard plots" in R:

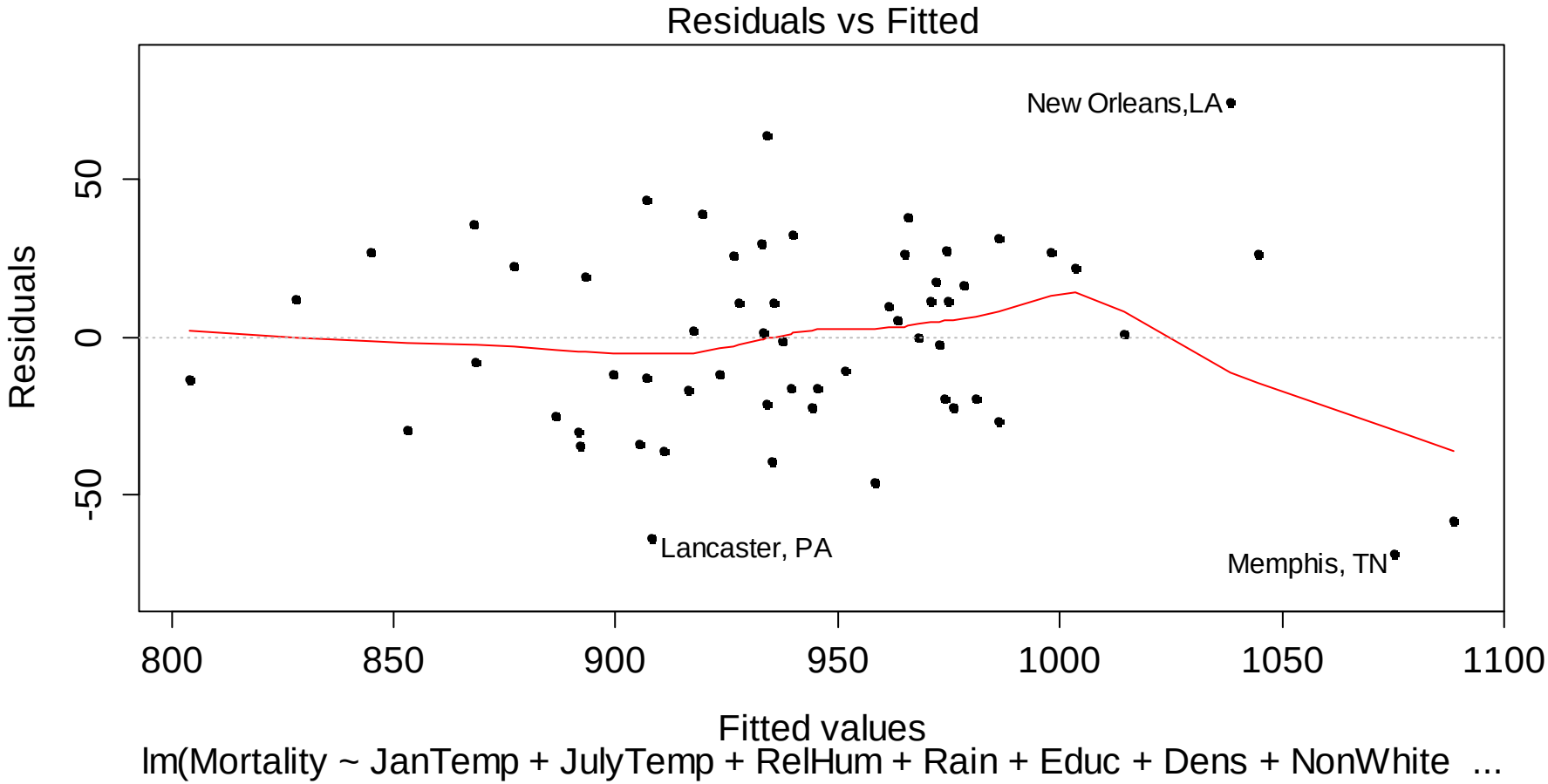
- Residuals vs. Fitted, i.e. Tukey-Anscombe-Plot
- Normal Plot (*uses standardized residuals*)
- Scale-Location-Plot (*uses standardized residuals*)
- Leverage-Plot (*uses standardized residuals*)

Some further tricks and ideas:

- Residuals vs. predictors
- Partial residual plots
- Residuals vs. other, arbitrary variables
- Important: Residuals vs. time/sequence

Applied Statistical Regression
AS 2014 – Multiple Regression

Tukey-Anscombe-Plot: Residuals vs. Fitted



Tukey-Anscombe-Plot: Residuals vs. Fitted

Some statements:

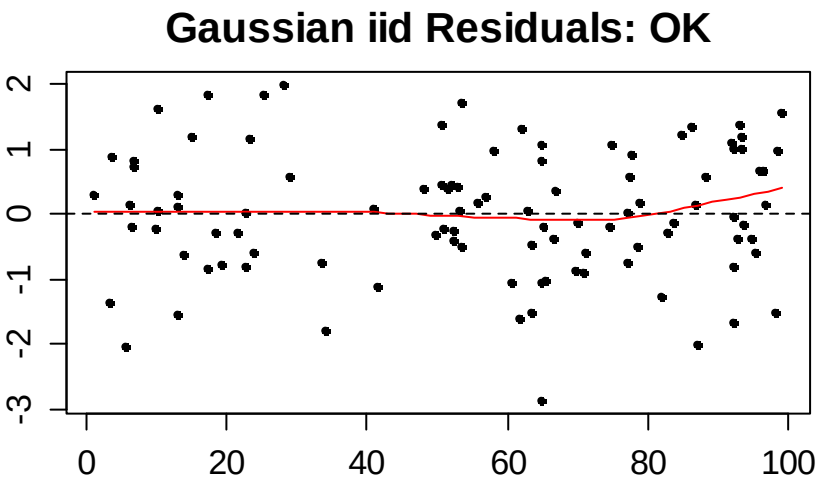
- is the most important residuals plot!
- is useful for finding structural model deficiencies $E[E_i] \neq 0$
- if $E[E_i] \neq 0$, the response/predictor relation might be nonlinear, or some important predictors/interactions may be missing.
- it is also possible to detect non-constant variance (→ then, the smoother does not deviate from 0)

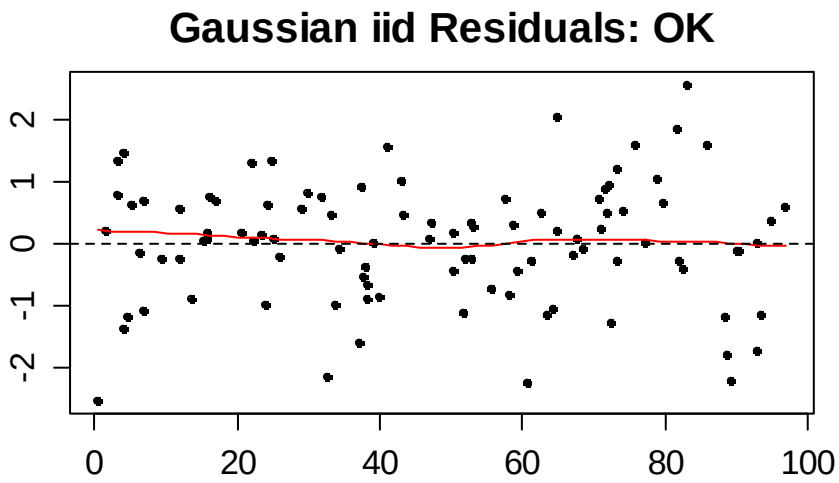
When is the plot OK?

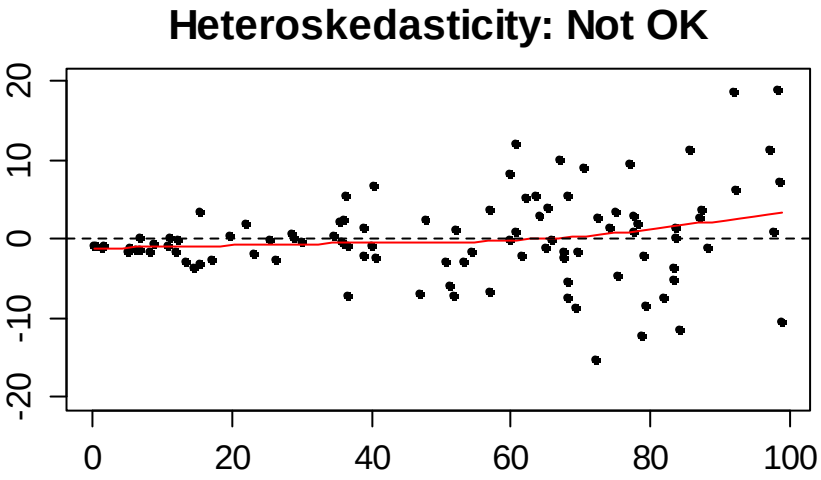
- the residuals scatter around the x-axis without any structure
- the smoother line is horizontal, with no systematic deviation
- there are no outliers

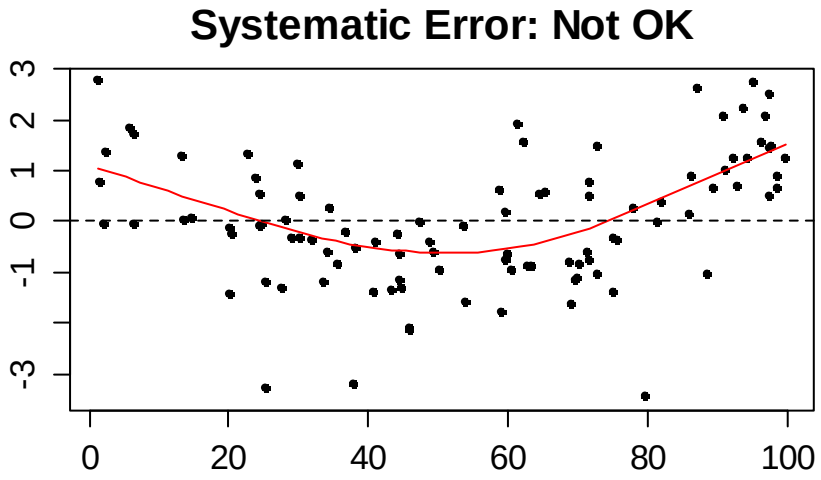
Applied Statistical Regression
AS 2014 – Multiple Regression

Tukey-Anscombe-Plot: Residuals vs. Fitted



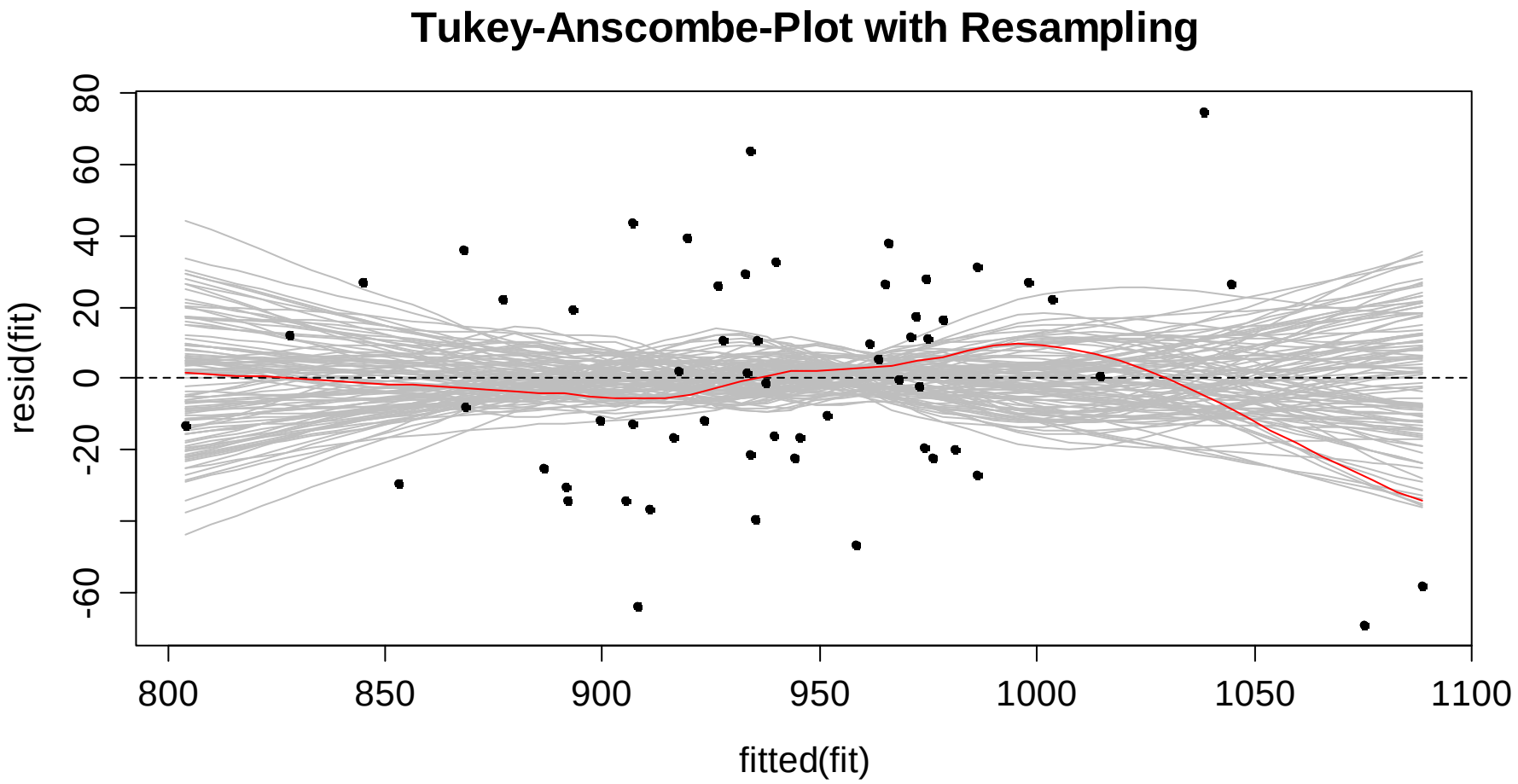






Applied Statistical Regression
AS 2014 – Multiple Regression

Tukey-Anscombe-Plot: Residuals vs. Fitted



Tukey-Anscombe-Plot

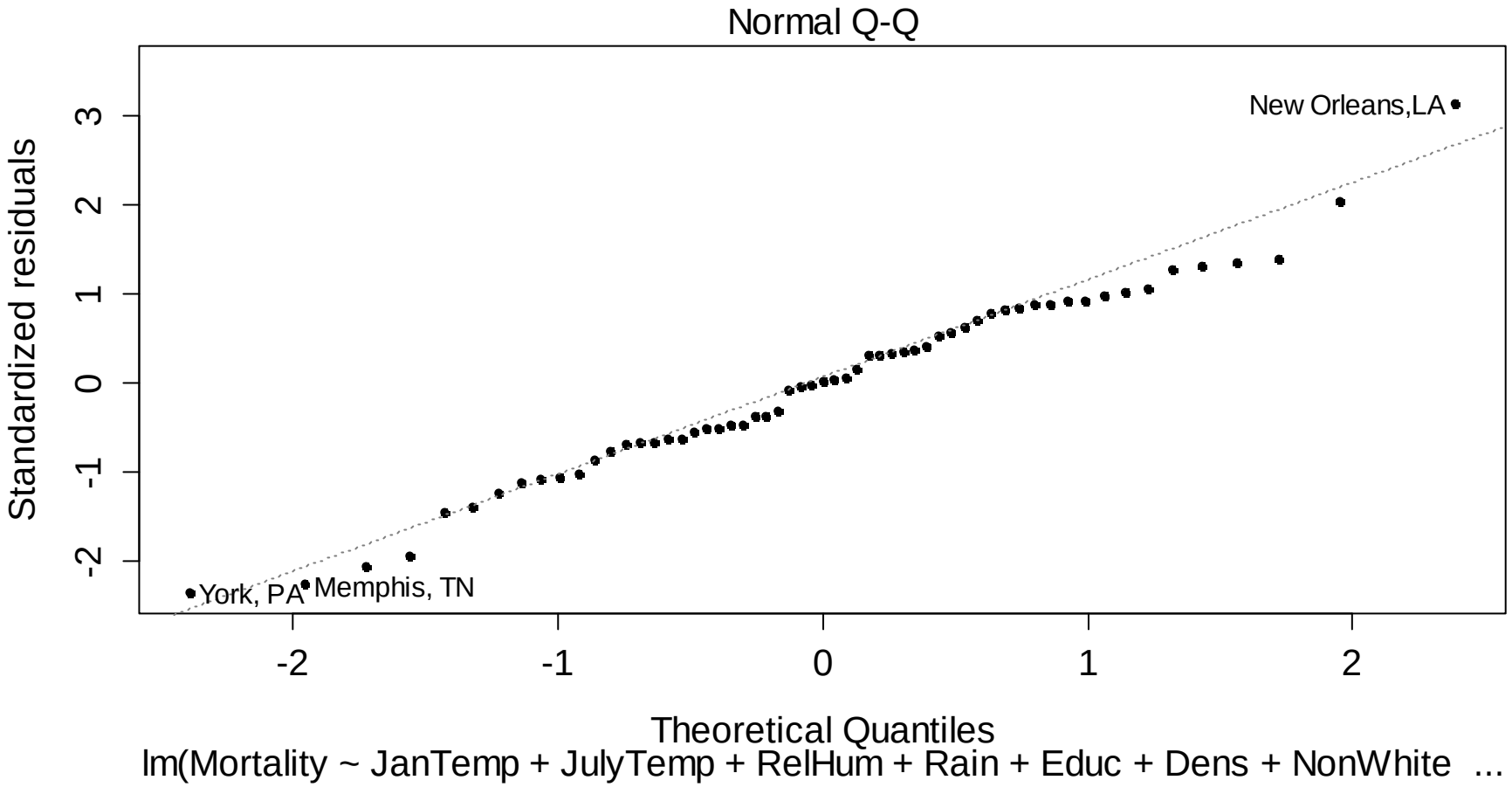
When the Tukey-Anscombe-Plot is not OK:

- If a systematic error is present, i.e. if the smoother deviates from the x-axis and hence $E[E_i] \neq 0$, it is mandatory to take some action. We recommend:
  - "*fit a better model*". In many cases, performing some log-transformations on the response and/or predictor(s) helps.
  - sometimes it also means that important predictors are missing. These can be completely novel variables, terms of higher order or interaction term.
- Non-constant variance: transformations usually help!

Applied Statistical Regression
AS 2014 – Multiple Regression

Normal Plot

Plot the residuals $\tilde{r}_i$ versus `qnorm(i/(n+1),0,1)`



Normal Plot

Is useful for:

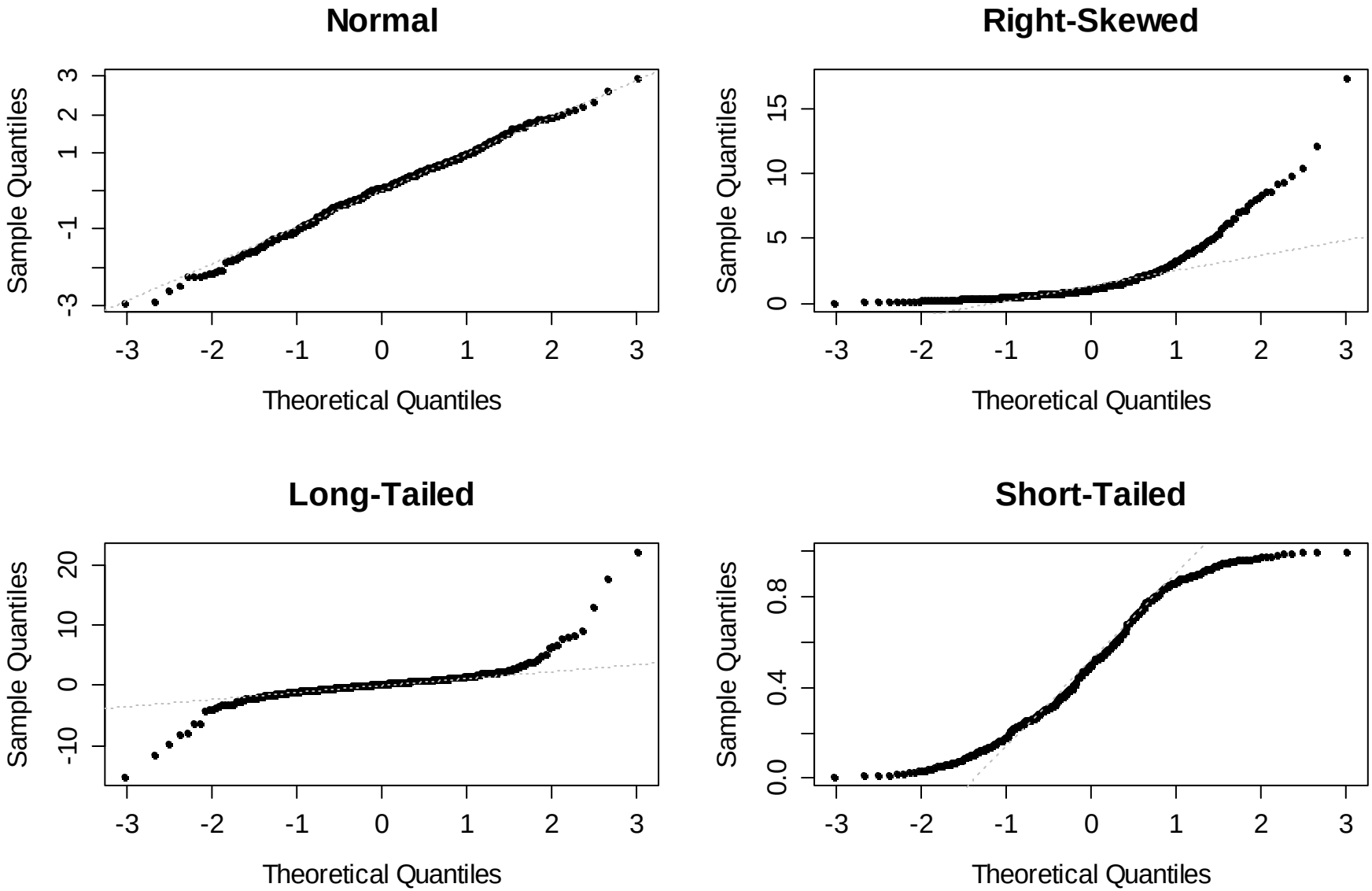
- for identifying non-Gaussian errors: $E_i \sim^{!} N(0, \sigma_E^2 I)$

When is the plot OK?

- the residuals $\tilde{r}_i$ must not show any systematic deviation from line which leads to the 1st and 3rd quartile.
- a few data points that are slightly "off the line" near the ends are always encountered and usually tolerable
- skewed residuals need correction: they usually tell that the model structure is not correct. Transformations may help.
- long-tailed, but symmetrical residuals are not optimal either, but often tolerable. Alternative: robust regression!

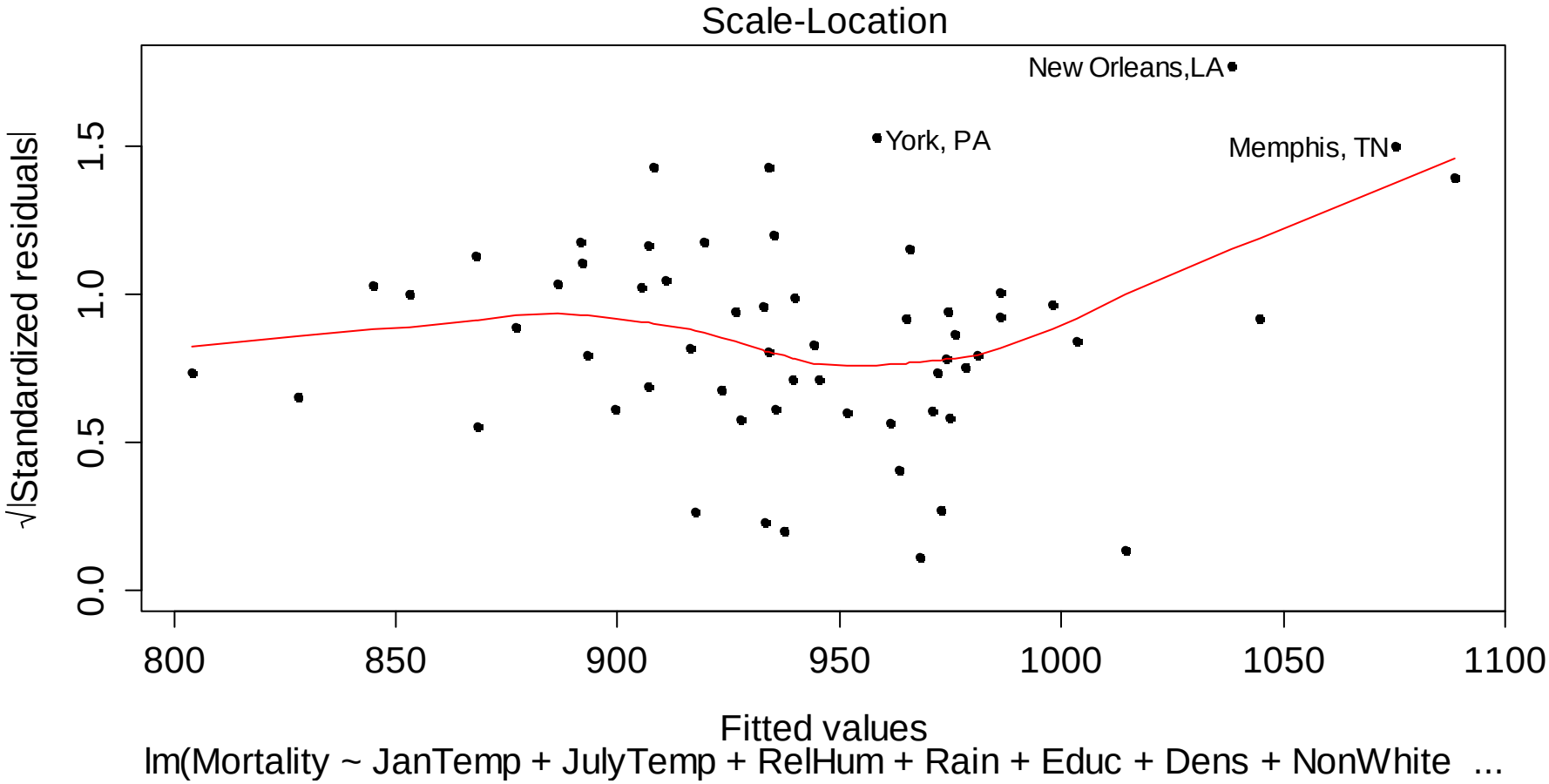
Applied Statistical Regression
AS 2014 – Multiple Regression

Normal Plot



Applied Statistical Regression
AS 2014 – Multiple Regression

Scale-Location-Plot: $\sqrt{|\tilde{r}_i|}$ *vs.* $\hat{y}_i$



Scale-Location-Plot

Is useful for:

- identifying non-constant variance: $Var(E_i) \neq \sigma_E^2$
- if that is the case, the model has structural deficencies, i.e. the fitted relation is not correct. Use a transformation!
- there are cases where we expect non-constant variance and do not want to use a transformation. This can the be tackled by applying weighted regression.

When is the plot OK?

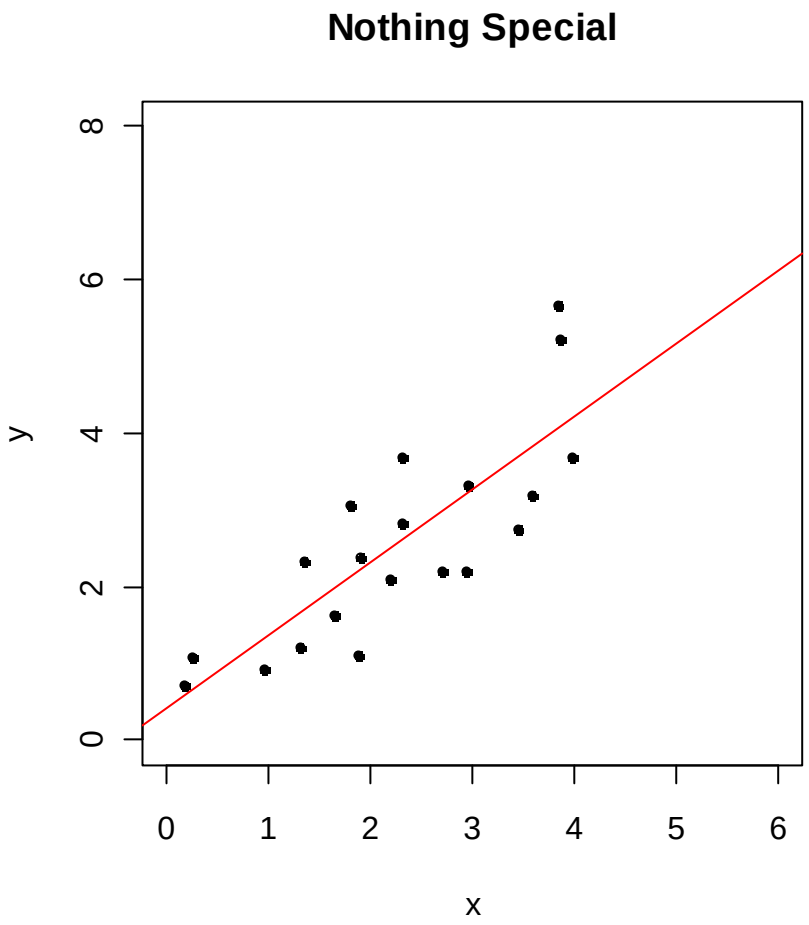
- the smoother line runs horizontally along the x-axis, without any systematic deviations.

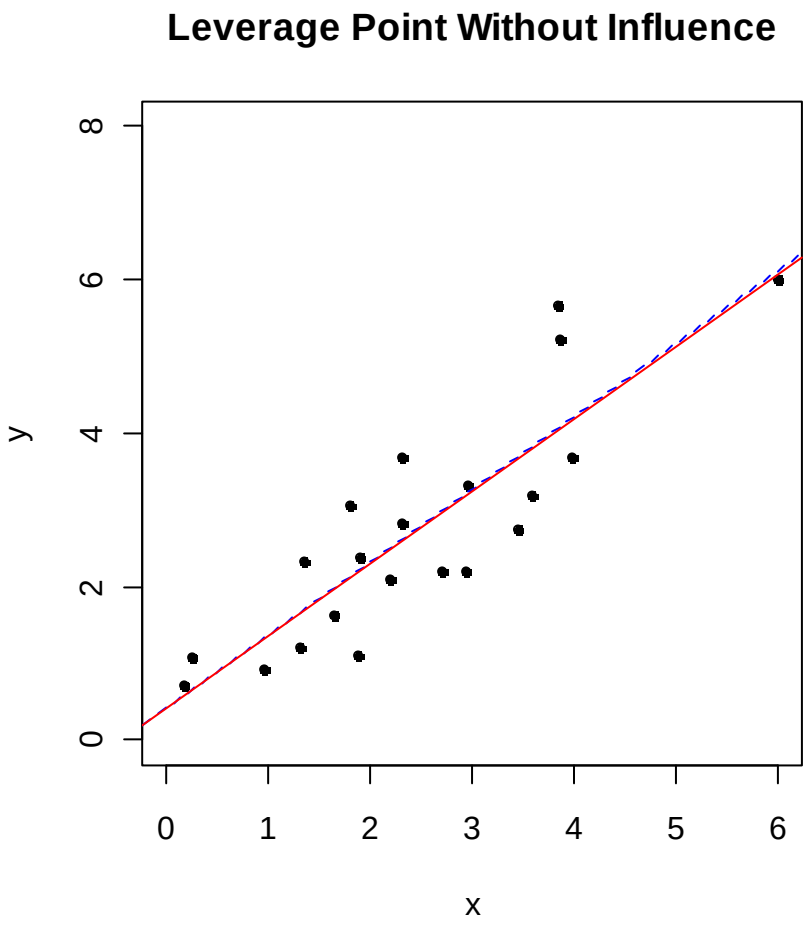
Unusual Observations

- There can be observations which do not fit well with a particular model. These are called ***outliers***.
- There can be data points which have strong impact on the fitting of the model. These are called ***influential observations***.
- A data point can fall under **none, one or both** the above definitions – there is no other option.
- A ***leverage point*** is an observation that lies at a "different spot" in predictor space. This is potentially dangerous, because it can have strong influence on the fit.

Applied Statistical Regression
AS 2014 – Multiple Regression

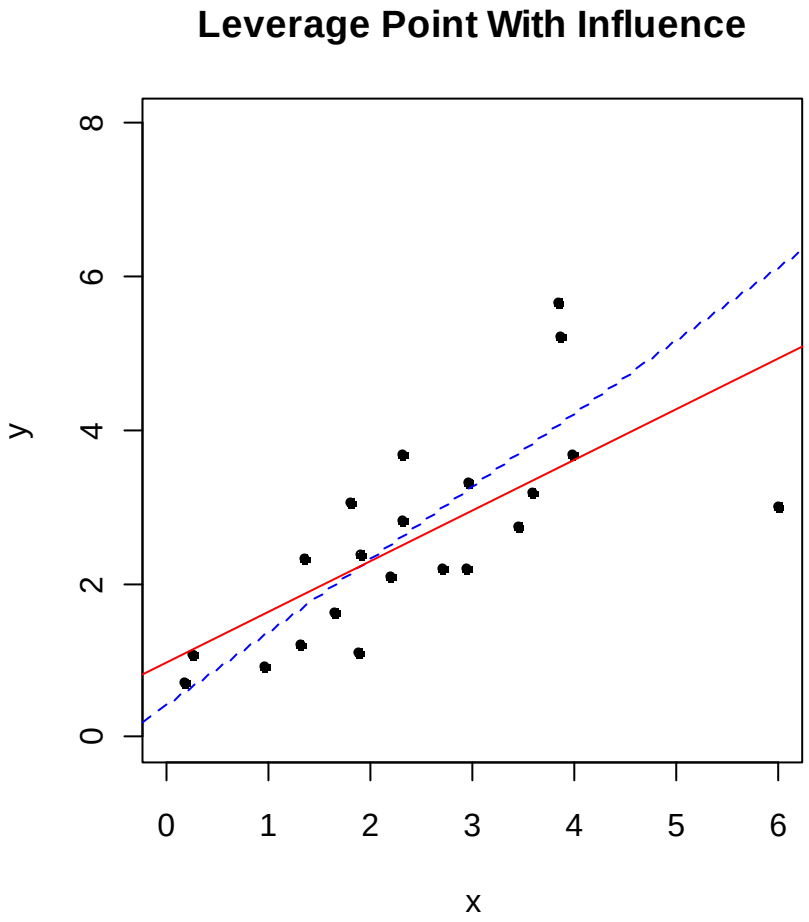
Unusual Observations

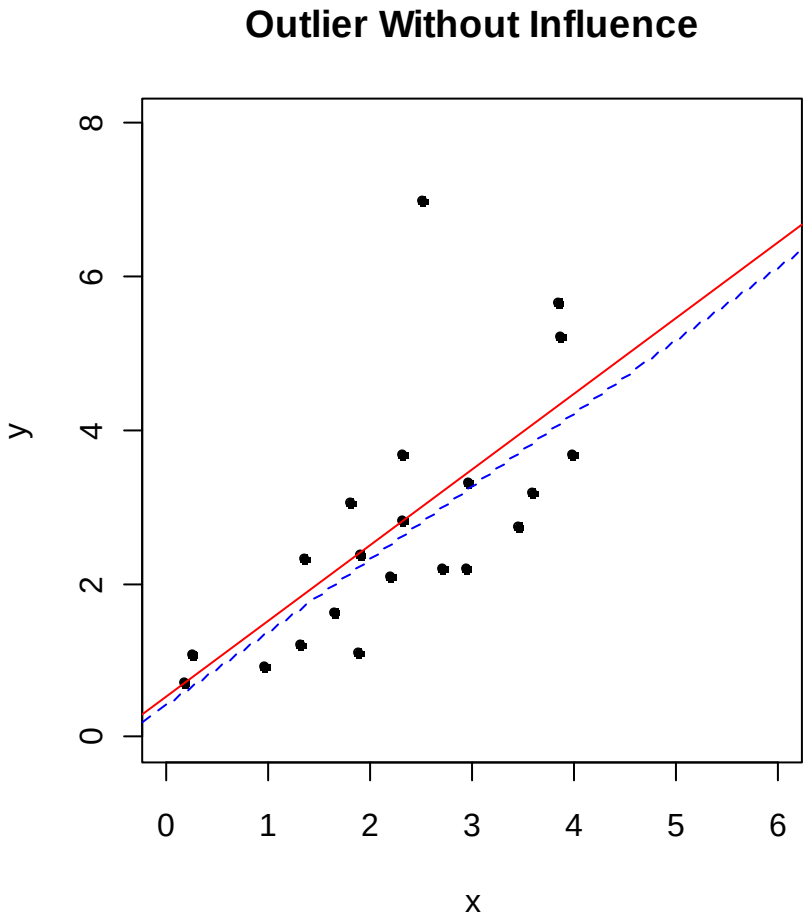




Applied Statistical Regression
AS 2014 – Multiple Regression

Unusual Observations





How to Find Unusual Observations?

1) Poor man‘s approach

Repeat the analysis n-times, where the i-th observation is left out. Then, the change is recorded.

2) Leverage

If y_i changes by Δy_i, then $h_{ii}\Delta y_i$ is the change in $\hat{y}_i$.
High leverage for a data point ($h_{ii} > 2(p+1)/n$) means that it forces the regression fit to adapt to it.

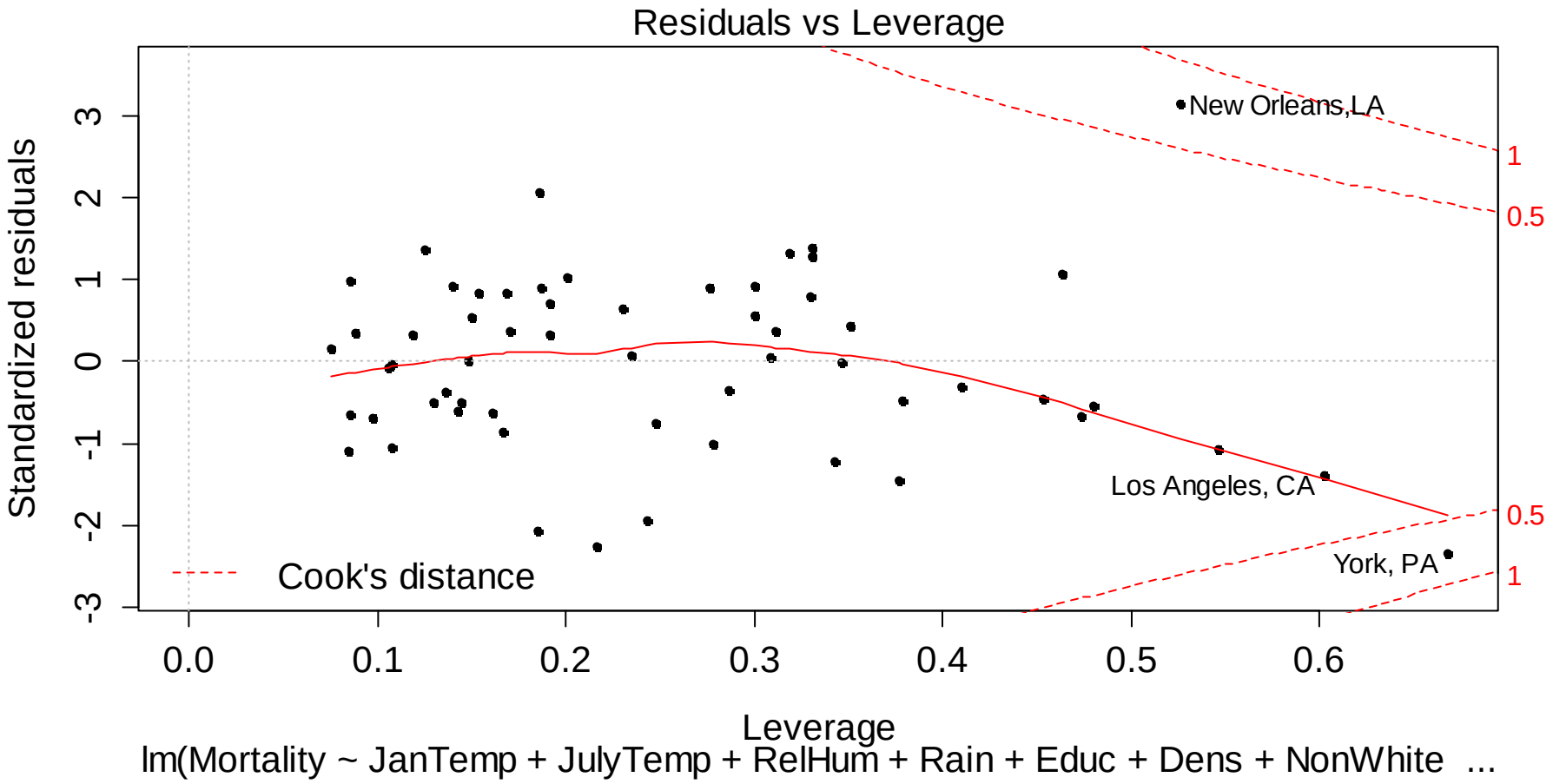
3) Cook‘s Distance

$$D_i = \frac{\sum(\hat{y}_j - y_{j(i)})^2}{(p+1)\sigma_E^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{(p+1)}$$

Be careful if Cook's Distance > 1.

Applied Statistical Regression
AS 2014 – Multiple Regression

Leverage-Plot: $\tilde{r}_i$ vs. Leverage h_{ii}



Leverage-Plot

Is useful for:

- identifying outliers, leverage points and influential observation at the same time.

When is the plot OK?

- no extreme outliers in y-direction, no matter where
- high leverage, here $h_{ii} > 2(p+1)/n = 2(4+1)/50 = 0.2$ is always potentially dangerous, especially if it is in conjunction with large residuals!
- This is visualized by the Cook's Distance lines in the plot: >0.5 requires attention, >1 requires much attention!

Leverage-Plot

What to do with unusual observations:

- First check the data for gross errors, misprints, typos, etc.
- Unusual observations are also often a problem if the input is not suitable, i.e. if predictors are extremely skewed, because first-aid-transformations were not done. Variable transformations often help in this situation.
- Simply omitting these data points is not a very good idea. Unusual observations are often very informative and tell much about the benefits and limits of a model.

Toolbox for Model Diagnostics

There are 4 "standard plots" in R:

- Residuals vs. Fitted, i.e. Tukey-Anscombe-Plot
- Normal Plot
- Scale-Location-Plot
- Leverage-Plot

Some further tricks and ideas:

- Residuals vs. predictors
- Partial residual plots
- Residuals vs. other, arbitrary variables
- Important: Residuals vs. time/sequence

More Residual Plots

General Remark:

We are allowed to plot the residuals versus any arbitrary variable we wish. This includes:

- predictors that were used
- potential predictors which were not (yet) used
- other variables, e.g. time/sequence of the observations

The rule is:

No matter what the residuals are plotted against, there must not be any non-random structure. Else, the model has some deficiencies, and needs improvement!

Example

Description of the Dataset:

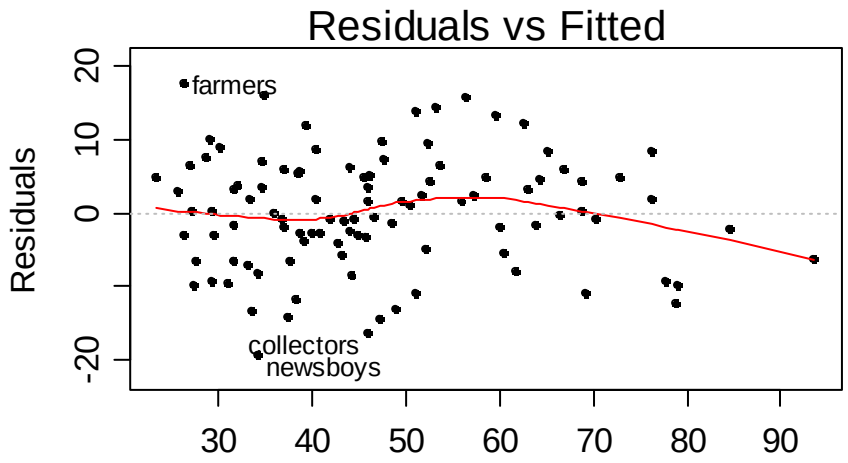
We are given a measurement of the prestige of 102 different profession. Moreover, we have 5 different variables that could be used as predictors. The data origin from Canada.

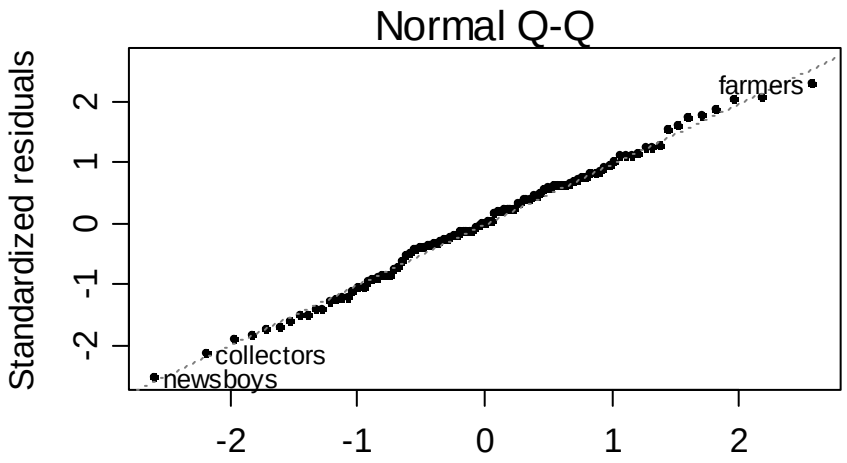
	educ	income	women	prest	cens	type
gov.administrator	13.11	12351	11.16	68.8	1113	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
accountants	12.77	9271	15.70	63.4	1171	prof

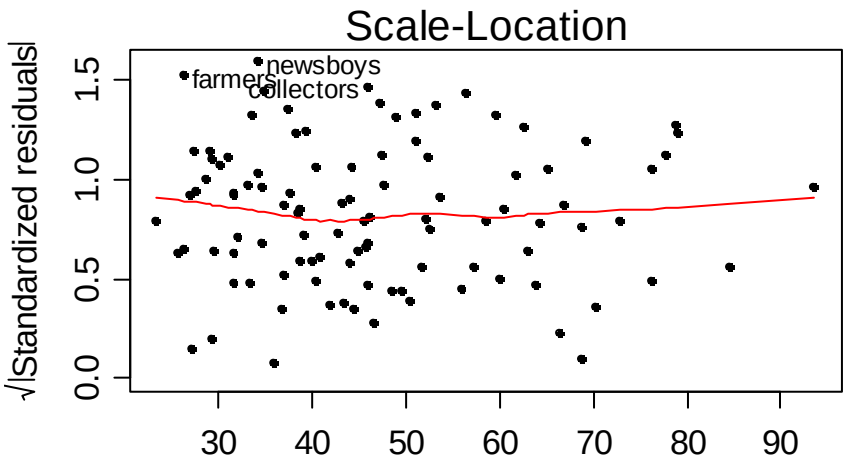
We start with fitting the model: `prestige ~ income + education`, the other three remaining (potential) predictors variables are first omitted in order the study the deficiencies in the model.

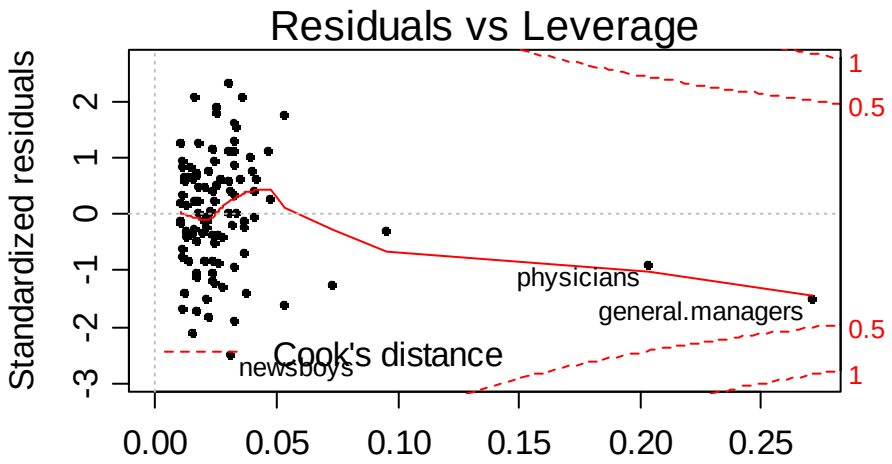
Applied Statistical Regression
AS 2014 – Multiple Regression

Standard Residual Plots with 2 Predictors



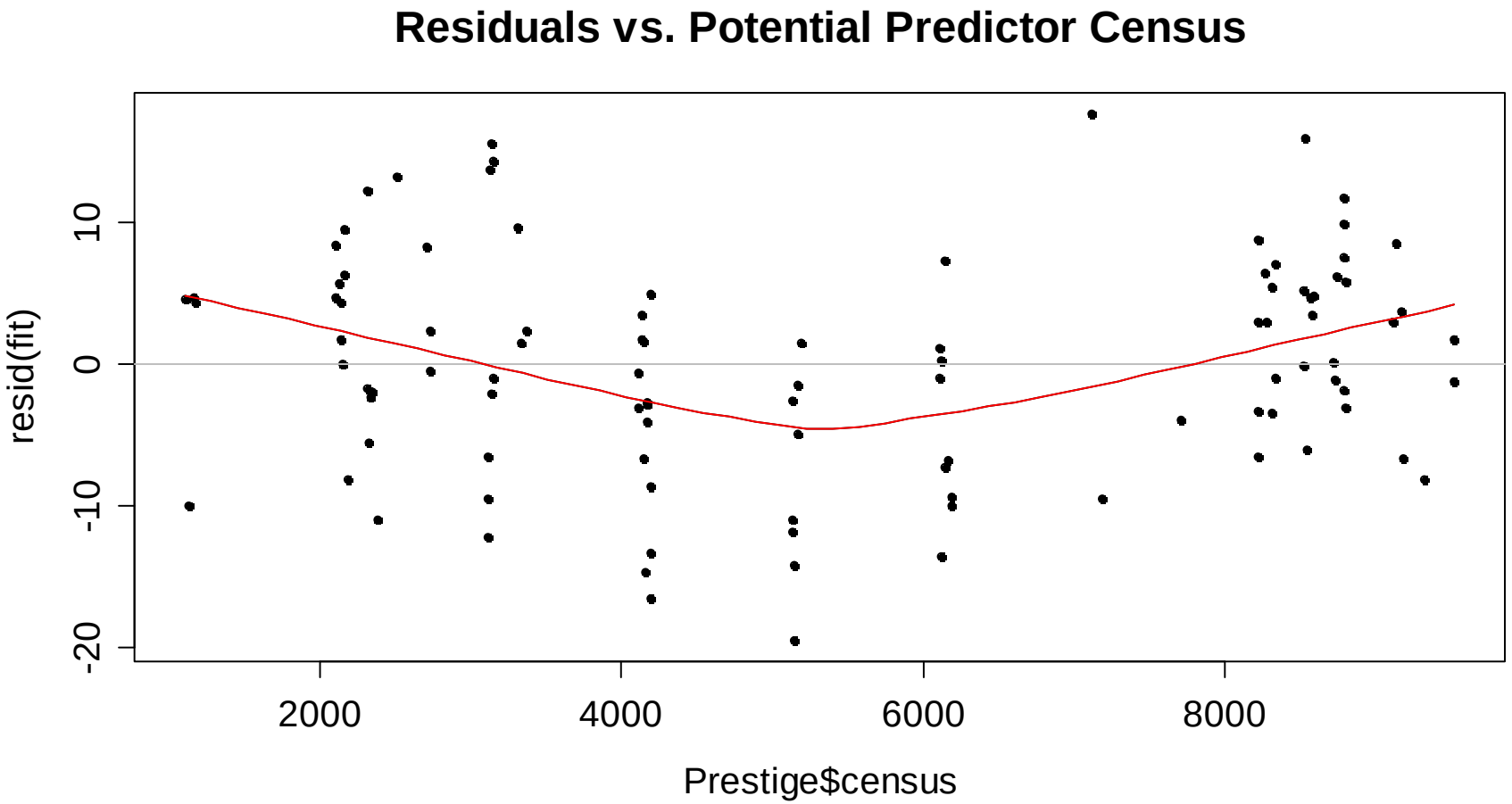






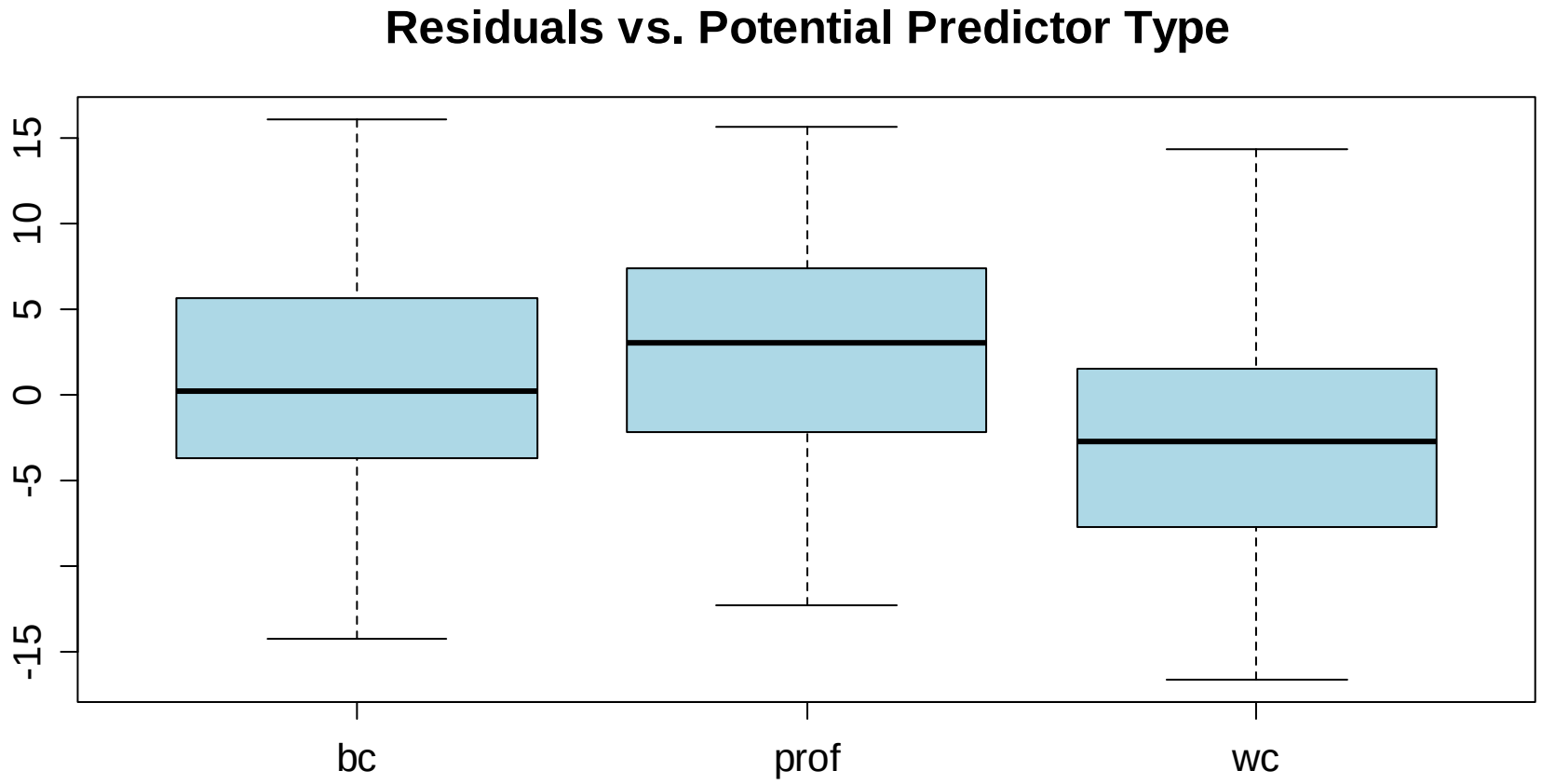
Applied Statistical Regression
AS 2014 – Multiple Regression

Residuals vs. Census



Applied Statistical Regression
AS 2014 – Multiple Regression

Residuals vs. Type



Motivation for Partial Residual Plots

Problem:

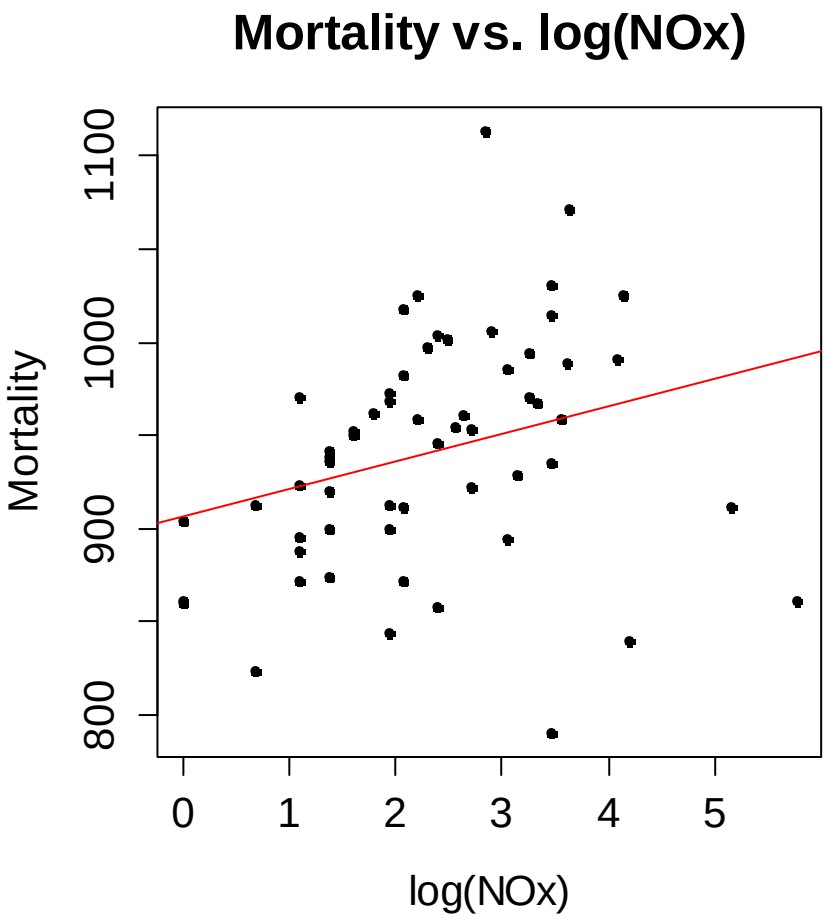
We sometimes want to learn about the relation between one predictor and the response, and also visualize it. Is it also of importance whether that relation is linear or not.

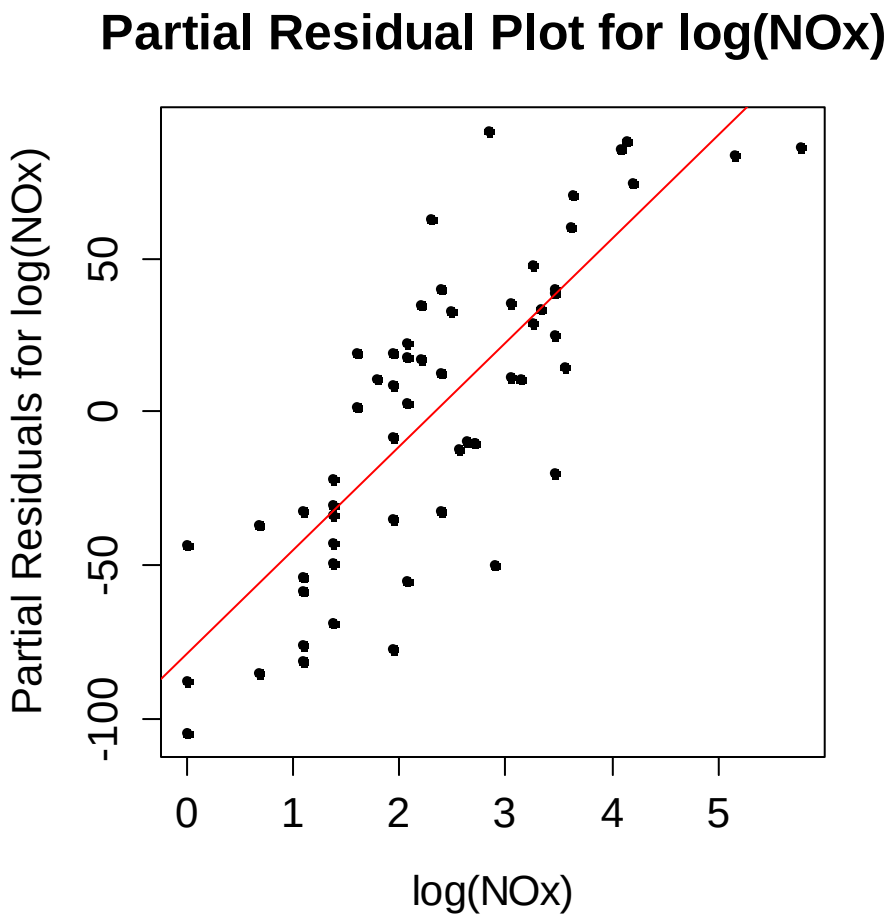
How can we infer this?

- the plot of response y vs. predictor x_k can be deceiving!
- the reason is that the other predictors $x_1, ..., x_{k-1}, x_{k+1}, ..., x_p$ also influence the response and thus blur our impression
- thus, we require a plot which only shows the "isolated“ influence of predictor x_k on the response y.

Applied Statistical Regression
AS 2014 – Multiple Regression

Partial Residual Plots: First Example





Applied Statistical Regression
AS 2014 – Multiple Regression

Partial Residual Plots

Idea:

We remove the estimated effect of all the other predictors from the response and plot this versus the predictor x_k.

$$y - \sum_{k \neq j} x_j \hat{\beta}_j = \hat{y} + r - \sum_{k \neq j} x_j \hat{\beta}_j = x_k \hat{\beta}_k + r$$

We then plot these so-called partial residuals versus the predictor x_k. We require the relation to be linear!

Partial residual plots in R:

- `library(car); crPlots(...)`
- `library(faraway); prplot(...)`

Partial Residual Plots: Example

We try to predict the prestige of a number of 102 different profession with a set of 2 predictors:

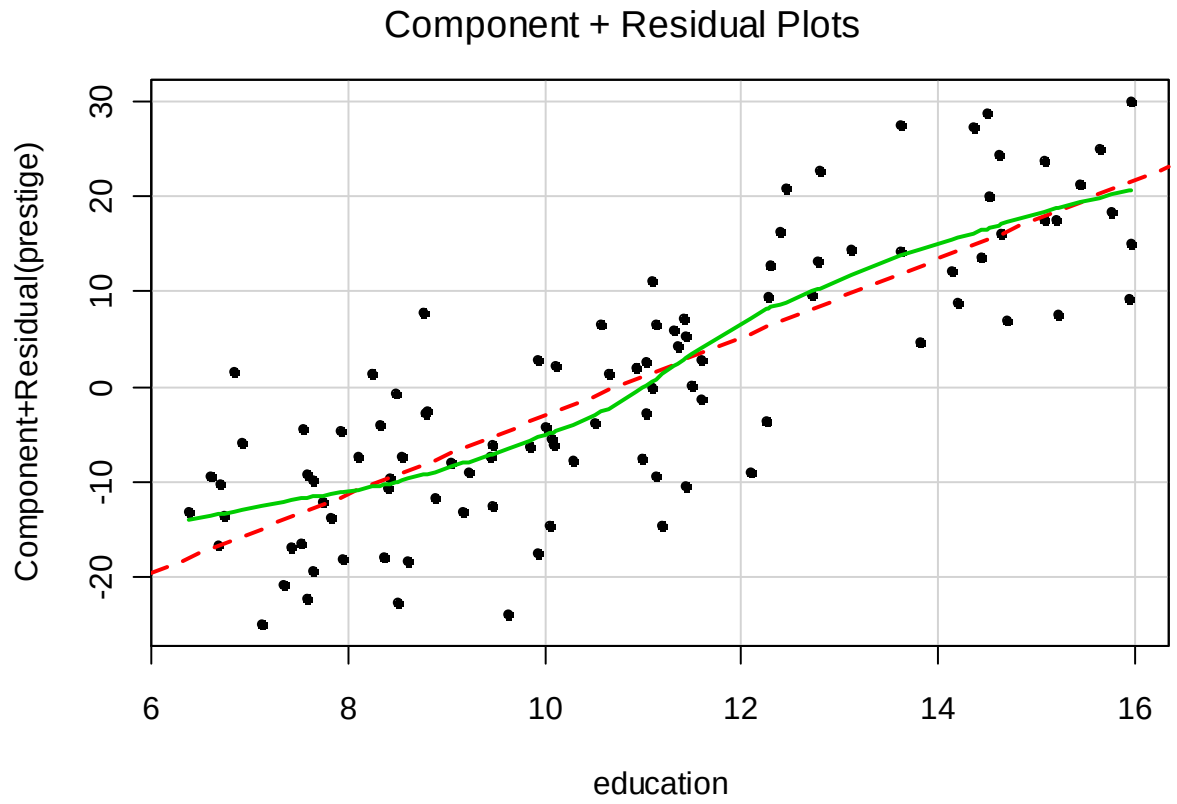
```
prestige ~ education + income
```

```
> data(Prestige)
> head(Prestige)
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
...
```

Applied Statistical Regression
AS 2014 – Multiple Regression

Partial Residual Plots: Education

```
> library(car); data(Prestige)
> fit <- lm(prestige ~ education + income, data=Prestige)
> crPlots(fit, layout=c(1,1))
```



For variable education, we seem to have made a reasonable choice:

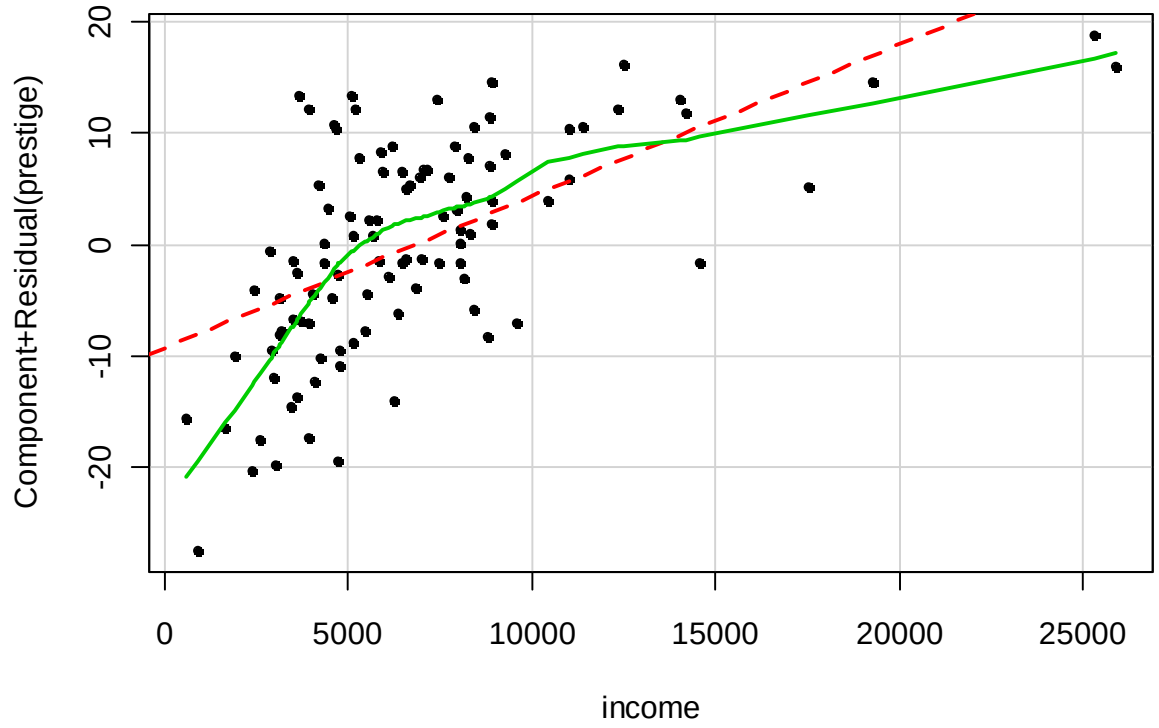
→ +/- linear relation

→ <12y vs. >12y ???

Applied Statistical Regression
AS 2014 – Multiple Regression

Partial Residual Plots: Example

```
> library(car); data(Prestige)
> fit <- lm(prestige ~ education + income, data=Prestige)
> crPlots(fit, layout=c(1,1))
```



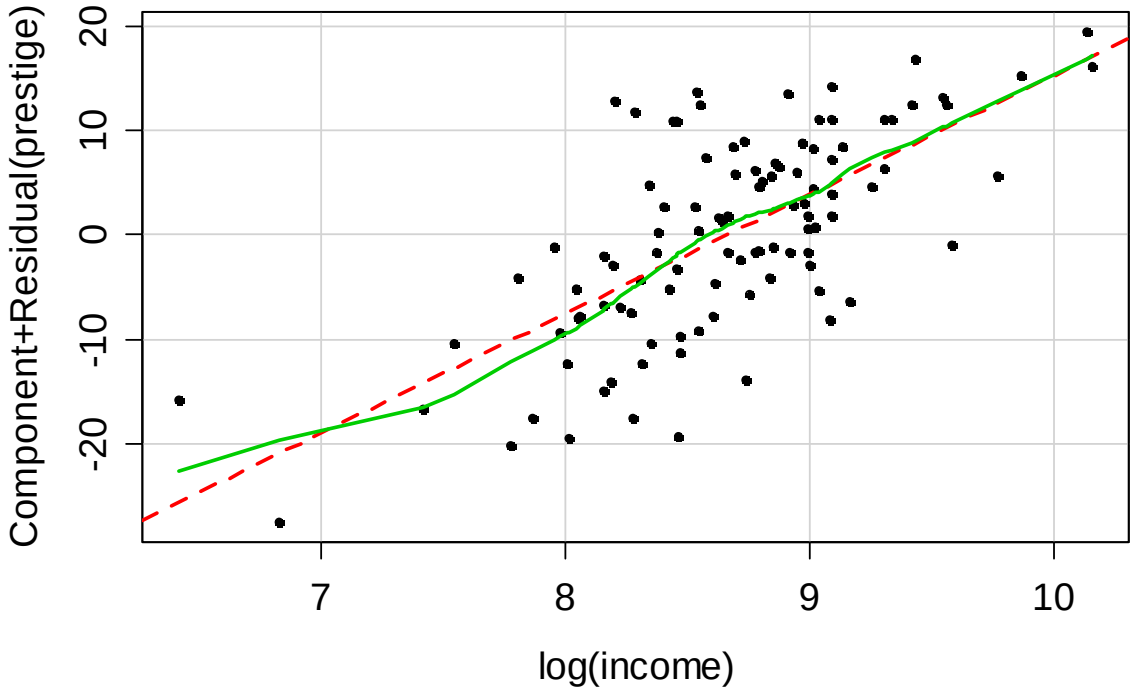
Evident non-linear influence of income on prestige.

→ not a good fit!
→ correction needed

Applied Statistical Regression
AS 2014 – Multiple Regression

Partial Residual Plots: Example

```
> library(car); data(Prestige)
> fit <- lm(prestige ~ education + log(income), Prestige)
> crPlots(fit, layout=c(1,1))
```



After a log-trsf of predictor 'income', things are fine

Partial Residual Plots

Summary:

Partial residual plots show the marginal relation between a predictor x_k and the response y.

When is the plot OK?

If the red line with the actual fit, and the green line of the smoother do not show systematic differences.

What to do if the plot is not OK?

- apply a transformation
- add further predictors into the model
- think about potential interaction terms

Checking for Correlated Errors

Background:

For LS-fitting we require uncorrelated errors. For data which have timely or spatial structure, this condition happens to be violated quite often.

Example:

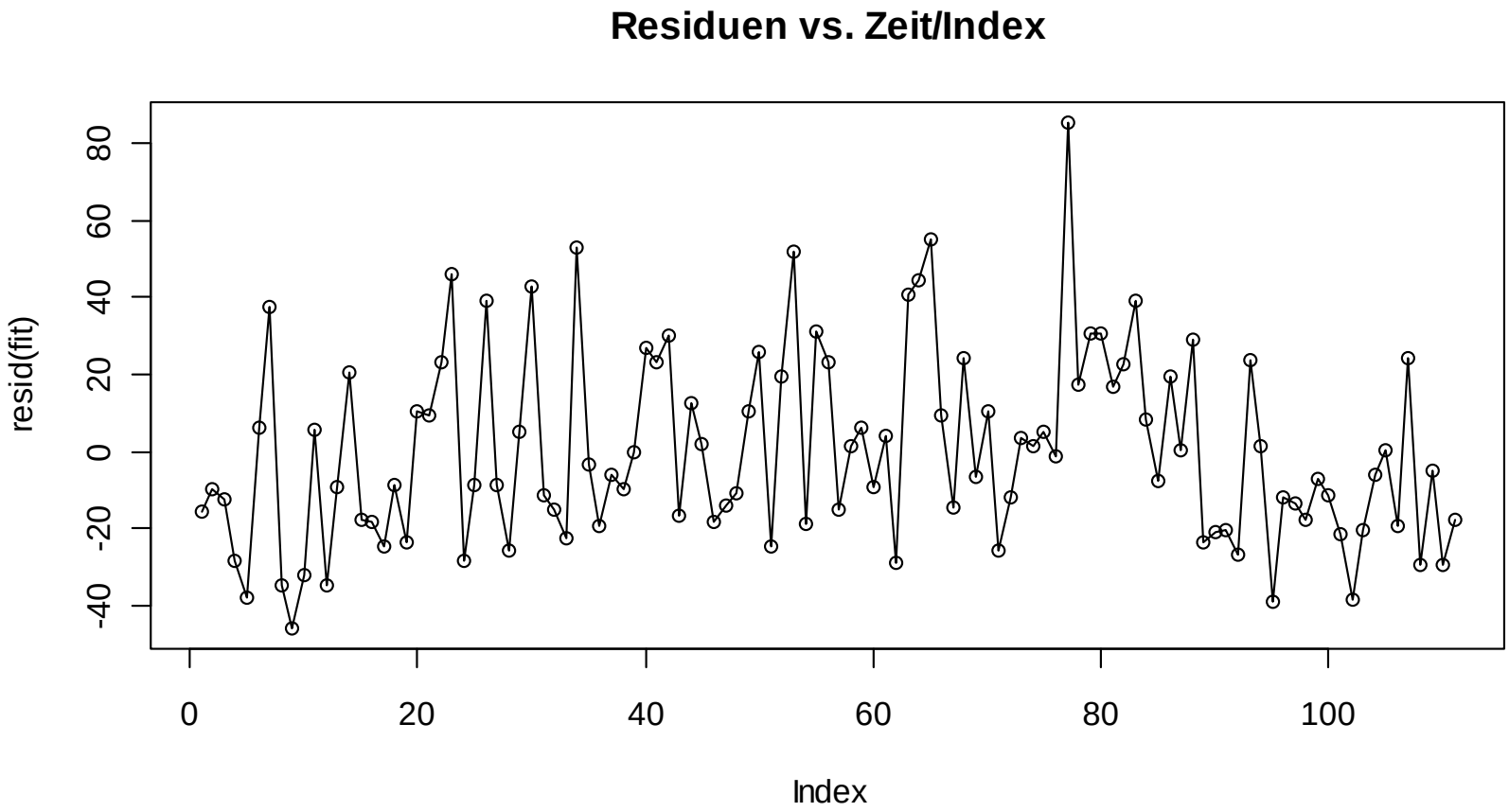
- `library(faraway); data(airquality)`
- `Ozone ~ Solar.R + Wind`
- Measurements from 153 consecutive days in New York
- data have a timely sequence

→ **to be handled with care!**

Applied Statistical Regression
AS 2014 – Multiple Regression

Residuals vs. Time/Index

```
> plot(resid(fit)); lines(resid(fit))
```



Alternative: Durbin-Watson-Test

The Durbin-Watson-Test checks if consecutive observations show a sequential correlation:

Test statistic: $$DW = \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}$$

- under the null hypothesis "no correlation", the test statistic has a χ^2- distribution. The p-value can be computed.
- the DW-test is somewhat problematic, because it will only detect simple correlation structure. When more complex dependency exists, it has very low power.

Applied Statistical Regression
AS 2014 – Multiple Regression

Durbin-Watson-Test

R-Hints:

```
> library(lmtest)
> dwtest(Ozone ~ Solar.R + Wind, data=airquality)
        Durbin-Watson test
data:  Ozone ~ Solar.R + Wind
DW = 1.6127, p-value = 0.01851
alternative hypothesis: true autocorrelation is greater than 0
```

The null hypothesis is rejected for the alternative that the true autocorrelation exceeds zero. From this we conclude that the residuals are not uncorrelated.

In the exercises, there is an deeper discussion of this problem...

Residuals vs. Time/Index

When is the plot OK?

- There is no systematic structure present
- There are no long sequences of pos./neg. residuals
- There is no back-and-forth between pos./neg. residuals
- The p-value in the Durbin-Watson test is >0.05

What to do if the plot is not OK?

1) Search for and add the "forgotten" predictors
2) Using the generalized least squares method (GLS)
   → to be discussed in *Applied Time Series Analysis*
3) Estimated coefficients and fitted values are not biased, but confidence intervals and tests are: be careful!

Weighted Regression

When to use?

Weighted regression is used when symmetrically distributed errors have zero expectation, but have non-constant variance. This violation might have been recognized by theoretical consideration (or more rarely) in the Scale-Location plot.

Important:

If non-constant variance is observed together with non-zero expectation of the error and/or skewed errors, then a trsf. of either response or some predictors is almost always better than using weighted regression.

Weighted Regression: Model

The model is:

$$y = X\beta + E \text{, where } E \sim N(0, \sigma_E^2 \Sigma)$$

→ For the non-weighted ordinary least squares regression, the error covariance matrix is the identity: $\Sigma = I$

→ We still assume uncorrelated errors, but no longer do we assume constant variance. The covariance matrix can thus be:

$$\Sigma = diag\left(\frac{1}{w_1}, \frac{1}{w_2}, ..., \frac{1}{w_n}\right) \neq I$$

Weighted Regression: And Now?

In a weighted least squares problem, the regression coefficients are estimated by minimizing a weighted sum of squares:

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

If the design matrix has full rank, this minimization problem has an explicit and unique solution. Moreover:

- Observations with small variance (i.e. where one is "sure" about the position of the data point) obtain large weight in the regression fit, and vice versa.

Where Are the Weights from?

1) If the response y_i is the mean from several independent observations, but not the same number of every data point. Then use: $w_i = n_i$.

   Example: Regression where daily cost in a mental hospital is explained with some socio-demographic predictors. The response variable is:

   "Total cost for the stay" / "Length of stay in days"

   The bigger the number of days that were used for assessing the cost, the more precise (=lower variance) the average cost is determined.

Where are the weights from?

2) One knows or can easily see that the variance in the residuals is proportional to a predictor.
   Then, we use: $w_i = 1/x_i$

   Example: **see Exercises...**

3) If non-constant variance is only "observed", but the cause is unknown (with respect to 1) and 2) above), the we can still try to first fit an ordinary least squares regression and use it for estimating weights, which will then be used in an weighted linear regression.

   Example: none...

Robust Regression

When to use?

Robust regression is used if the residuals are symmetrically distributed and have expectation zero, but are more heavy-tailed than the Gaussian distribution suggests.

Be careful:

If long-tailed resdiuals appear in conjunction with a non-idle Tukey-Anscombe-Plot, and/or with non-constant variance, or if the residuals are skewed, then applying transformations is more appropriated than using robust regression.

Also if there are a few gross outliers, it's better to study these in detail, rather than just applying robust regression.

Robust Regression: Model

The model in robust regression is:

$$y = X\beta + E \text{, where } E \sim^{!} N(0, \sigma_E^2 \Sigma)$$

→ The errors are assumed to be symmetrically distributed, but more heavy-tailed than the Gaussian.

→ In this case, the LS-method is no longer optimal/efficient. There are better estimators for the regression coefficients.

→ Short-tailed errors do not need special attention. In such cases, it is fine to apply the ordinary LS method.

Robust Regression: Idea

In robust regression, observations with large residuals obtain a smaller weight. This is implemented by using a modified "loss function", i.e. no longer the LS-criterion, that measures the quality of the fit:

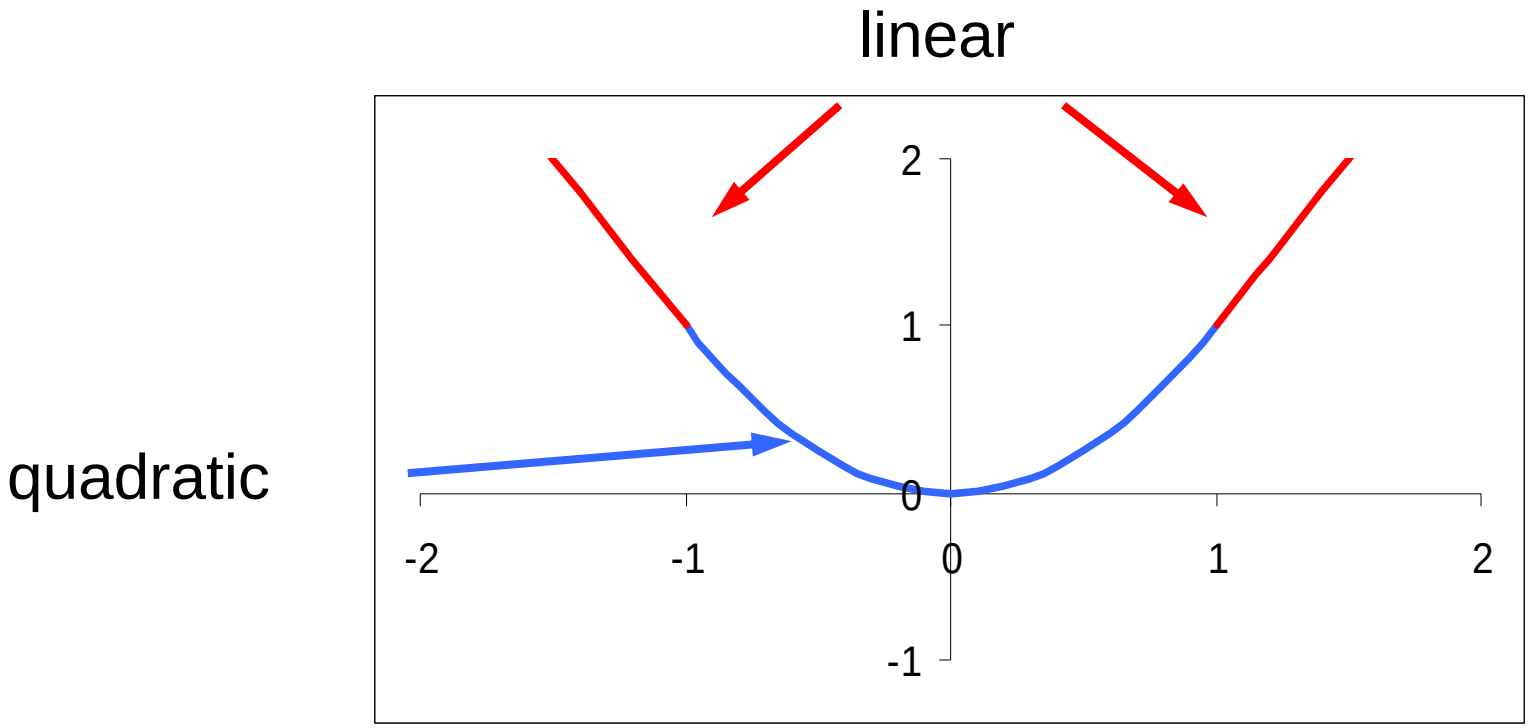
$$\sum_{i=1}^{n} \rho(r_i), \text{ where } \rho(x) = \begin{cases} x^2/2 & \text{if } |x| \le c \\ c|x| - c^2/2 & \text{if } |x| > c \end{cases}$$

Visualization: see next slide!

There is no solution which can be written in closed form, and an optimization procedure needs to be employed. This is done by solving iteratively reweighted least squares regressions.

Huber Loss Function

This function is used as the default in R-function `rlm()` from `library(MASS)`. There are many other suggestions…



Applied Statistical Regression
AS 2014 – Multiple Regression

Robust Regression: R-Code

```
> library(MASS)
> fit.rlm <- rlm(Mortality ~ JanTemp + … + log(SO2), data=…)
```

→ This uses the Huber loss function

→ The summary is different!

```
summary(fit.rlm)
Coefficients: Value Std. Error    t value
(Intercept) 945.4414  251.6184     3.7574
JanTemp      -1.2313   0.6788     -1.8139
log(SO2)     13.0484   4.6444      2.8095
---
Residual standard error: 30.17 on 46 degrees of freedom
```

Collinearity = Correlated Predictors

If ≥ 2 predictors are strongly correlated, i.e. explain very similar aspects of the response, OLS estimation is difficult. The regression coefficients will be less precise, and the interpretation of the results is more difficult.

There is a need to recognize collinearity!

1) *Plot the correlation matrix of the predictors*

```
plotcorr(cor(my.dat))
```

2) *Variance Inflation Factors*

$$Var(\hat{\beta}_k) = \sigma_E^2 \cdot \frac{1}{1-R_k^2} \cdot \frac{1}{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)}$$

Applied Statistical Regression
AS 2014 – Multiple Regression

How to Deal with Correlated Predictors?

1) **Generate new variables**

→ **see example on next slides...**

2) **Variable selection**

Only use the relevant predictors, and omit the redundant ones. This often helps a lot. We will be discussing variable selection in detail.

3) **The Lasso and Ridge Regression**

These are penalized OLS regression methods, which sparsely spend degrees of freedom. To be discussed later.

Example

Understanding how car drivers adjust their seat would greatly help engineers to design better cars. Thus, the measured

hipcenter = horizontal distance of hips to steering wheel

and tried to explain it with several predictors, namely:

Age	age in years
Weight	weight in pounds
HtShoes, Ht, Seated	height w/o, w/ shoes, seated height
Arm, Thigh, Leg	arm, thigh and leg length

We first fit a model with all these (correlated!) predictors

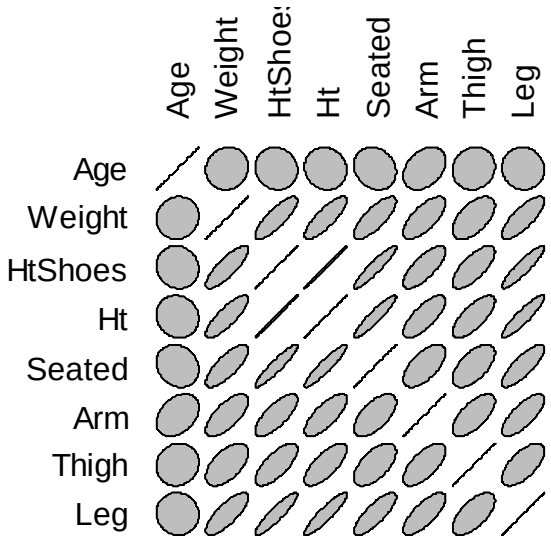
Example: Fit with All Predictors

```
> library(faraway); data(seatpos)
> summary(lm(hipcenter~., data=seatpos))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared: 0.6866,Adjusted R-squared: 0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Applied Statistical Regression
AS 2014 – Multiple Regression

Collinearity in the Seat Position Example

```
> vif(fit)
      Age      Weight    HtShoes          Ht
 1.997931    3.647030 307.429378 333.137832
   Seated         Arm       Thigh         Leg
 8.951054    4.496368    2.762886    6.694291
```



VIF ≥ 5 is critical, VIF ≥ 10 is dangerous. The observed values mean that the standard errors of the estimates are inflated by a factor up to about 18x.

Example: Generating New Variables

The body height is certainly a key predictors when it comes to the position of the driver seat. We leave this as it was, and change several of the other predictors:

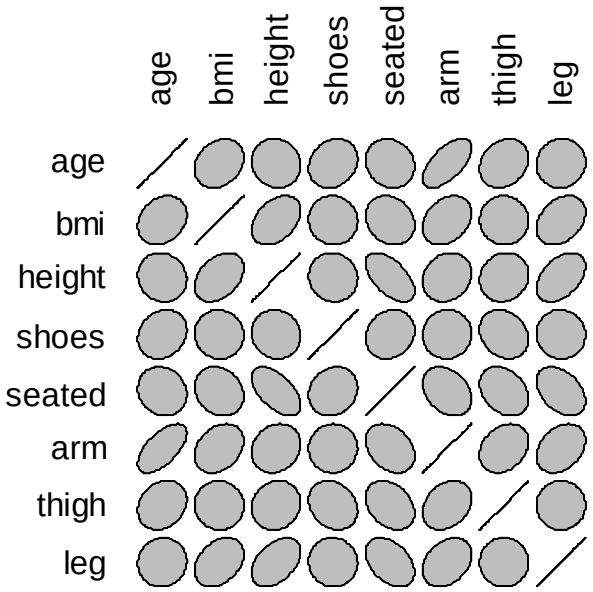
```
age    <- Age
bmi    <- (Weight*0.454)/(Ht/100)^2
shoes  <- HtShoes-Ht
seated <- Seated/Ht
arm    <- Arm/Ht
thigh  <- Thigh/Ht
leg    <- Leg/Ht
```

Does this solve the correlation problem...?

Applied Statistical Regression
AS 2014 – Multiple Regression

Example: New Correlation Matrix

```
> vif(fit00)
     age      bmi   height    shoes   seated
1.994473 1.408055 1.968447 1.155285 1.851884
     arm    thigh      leg
2.044727 1.284893 1.480397
```



Example: Fit with New Predictors

```
> summary(lm(hipc~., data=new.seatpos))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -632.0063   490.0451  -1.290    0.207
age           -0.7402     0.5697  -1.299    0.204
bmi           -0.4234     2.2622  -0.187    0.853
height         3.6521     0.7785   4.691 5.98e-05 ***
shoes          2.6964     9.8030   0.275    0.785
seated      -171.9495   631.3719  -0.272    0.787
arm          180.7123   655.9536   0.275    0.785
thigh        141.2007   443.8337   0.318    0.753
leg         1090.0111   806.1577   1.352    0.187

Residual standard error: 37.71 on 29 degrees of freedom
Multiple R-squared: 0.6867, Adjusted R-squared: 0.6002
F-statistic: 7.944 on 8 and 29 DF,  p-value: 1.3e-05
```

Applied Statistical Regression
AS 2014 – Multiple Regression

Variable Selection: Why?

We want to fit a regression model…

Case 1: functional form and predictors exactly known
→ *estimation, test, confidence and prediction intervals*

Case 2: neither functional form nor the predictors are known
→ *explorative model search among potential predictors*

Case 3: we are interested in only 1 predictor, but want to correct for the effect of other covariates
→ *which covariates we need to correct for?*

Question in cases 2 & 3: WHICH PREDICTORS TO USE?

Variable Selection: Technical Aspects

We want to keep a model small, because of

1) **Simplicity**
   → *among several explanations, the simplest is the best*

2) **Noise Reduction**
   → *unnecessary predictors leads to less accuracy*

3) **Collinearity**
   → *removing excess predictors facilitates interpretation*

4) **Prediction**
   → *less variables, less effort for data collection*

Method or Process?

- **Variable selection is not a method!** The search for the best predictor set is an iterative process. It also involves *estimation*, *inference* and *model diagnostics*.
- For example, outliers and influential data points will not only change a particular model – they can even have an impact on the model we select. Also variable transformations will have an impact on the model that is selected.
- Some iteration and experimentation is often necessary for variable selection. *The ultimate aim is finding a model that is smaller, but as good or even better than the original one.*

Example: Mortality Data

```
> summary(fit.orig)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1496.4915   572.7205   2.613  0.01224 *
JanTemp       -2.4479     0.8808  -2.779  0.00798 **
...
Dens          11.9490    16.1836   0.738  0.46423
NonWhite     326.6757    62.9092   5.193 5.09e-06 ***
WhiteCollar -146.3477   112.5510  -1.300  0.20028
...
---
Residual standard error: 34.23 on 44 degrees of freedom
Multiple R-squared: 0.7719, Adjusted R-squared: 0.6994
F-statistic: 10.64 on 14 and 44 DF,  p-value: 6.508e-10
```

Note: due to space constraints, this is only part of the output.

Backward Elimination with p-Values

Aim: Reducing the regression model such that the remaining predictors show a significant relation to the response.

How: We start with the full model and then exclude the least significant predictor in a step-by-step manner, as long as its p-value is greater than $\alpha_{crit} = 0.05$.

In R:
```
> fit <- update(fit, . ~ . - RelHum)
> summary(fit)
```

→ *Re-fit the model after each exclusion!*

→ *Wording:* **Backward Elimination with p-Values**

→ For prediction, one also uses $\alpha_{crit} = 0.10 / 0.15 / 0.20$

Applied Statistical Regression
AS 2014 – Multiple Regression

Example: Final Result

```
> ft09 <- update(ft08, .~.-WhiteCollar); summary(ft09)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 992.2069    79.6994  12.449  < 2e-16 ***
JanTemp      -2.1304     0.5017  -4.246 8.80e-05 ***
Rain          1.8122     0.5066   3.577 0.000752 ***
Educ        -16.4207     6.1202  -2.683 0.009710 **
NonWhite    268.2564    38.8832   6.899 6.56e-09 ***
NOx          18.3230     4.3960   4.168 0.000114 ***
---
Residual standard error: 33.47 on 53 degrees of freedom
Multiple R-squared: 0.7373, Adjusted R-squared: 0.7125
F-statistic: 29.75 on 5 and 53 DF,  p-value: 2.931e-14
```

→ 9 predictors are eliminated, 5 remain in the final model.

Interpretation of the Result

- The remaining predictors are now “more significant” than before. This is almost always the case. Do not overestimate the importance of these predictors!
- Collinearity among the predictors is usually at the root of this observation. The predictive power is first spread out among several predictors, then it becomes concentrated.
- **Important**: the removed variables can still be related to the response. If we run a simple linear regression, they can even be significant. In the multiple linear model however, there are other, better, more informative predictors.

Alternatives to Backward Elimination

Backward elimination that is based on p-values requires laborious handwork (*in R*) and has a few disadvantages...

- When the principal goal is prediction, then the resulting models are often too small, i.e. there are bigger models which yield a more accurate prognosis.
- From a (theoretical) mathematical viewpoint variable selection via the AIC/BIC criterions is more suitable.
- In a step-by-step backward elimination, the best model is often missed. Evaluating more models can be very beneficial for finding *the best one*...

The AIC/BIC Criteria

Aim: Judging the quality of a regression model

→ *Gauging Goodness-of-Fit vs. The Number of Predictors*

AIC-Criterion:

$$\begin{aligned} AIC &= -2\max(\log likelihood) + 2p \\ &= const + n\log(RSS/n) + 2p \end{aligned}$$

BIC-Criterion:

$$\begin{aligned} BIC &= -2\max(\log likelihood) + p\log n \\ &= const + n\log(RSS/n) + p\log n \end{aligned}$$

Backward Elimination with AIC/BIC

Aim: Reducing the regression model such that the remaining predictors are *necessary* for describing the response.

How: We start with the full model and then in a step-by-step manner exclude the predictor that leads to the biggest improvement in AIC/BIC.

In R:
```
> fit.aic <- step(fit, dir="backward", k=2)
> fit.bic <- step(fit, dir="backward", k=log(59))
```

→ *The variable selection stops when AIC/BIC cannot be improved anymore. There is neither a need nor a guarantee that the selected predictors are significant.*

Applied Statistical Regression
AS 2014 – Multiple Regression

Example: Models with AIC/BIC

AIC:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1035.5384	85.1924	12.155	< 2e-16	***
JanTemp	-2.0188	0.5043	-4.003	0.000200	***
Rain	1.9637	0.5146	3.816	0.000363	***
Educ	-11.7708	6.9613	-1.691	0.096842	.
NonWhite	261.5379	38.8830	6.726	1.35e-08	***
WhiteCollar	-139.2913	102.0379	-1.365	0.178101	
NOx	19.4440	4.4372	4.382	5.73e-05	***

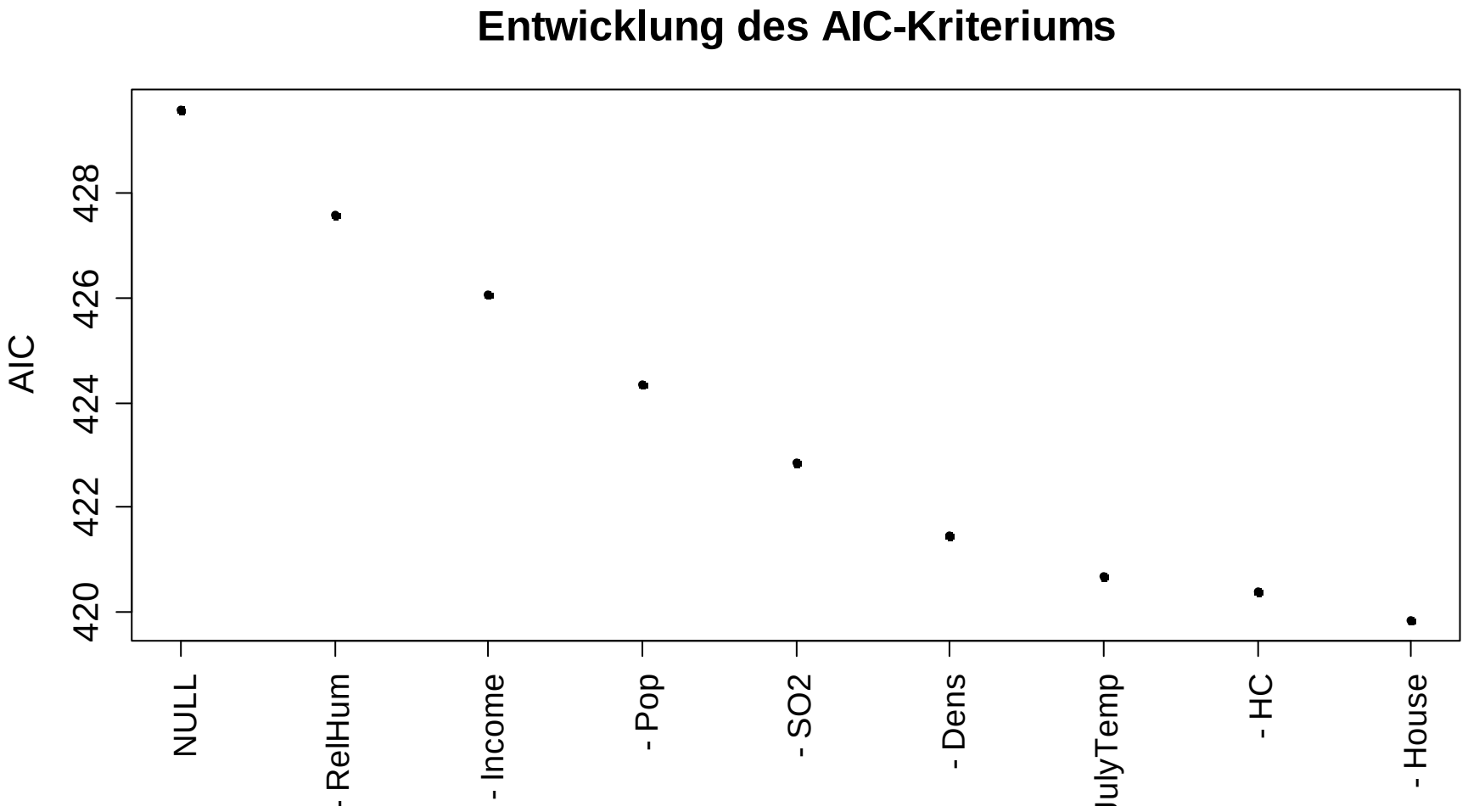
BIC:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	992.2069	79.6994	12.449	< 2e-16	***
JanTemp	-2.1304	0.5017	-4.246	8.80e-05	***
Rain	1.8122	0.5066	3.577	0.000752	***
Educ	-16.4207	6.1202	-2.683	0.009710	**
NonWhite	268.2564	38.8832	6.899	6.56e-09	***
NOx	18.3230	4.3960	4.168	0.000114	***

Applied Statistical Regression
AS 2014 – Multiple Regression

Visualization of Variable Selection

```
> plot(fit.aic$anova$AIC, ...)
```



AIC or BIC?

Usually, both criteria lead to similar models. BIC penalizes bigger models harder, with factor $\log n$ instead of factor 2.

→ *"BIC models" tend to be smaller than "AIC models"!*

Rule of the thumb for criterion choice:

- **BIC** is used when we are after a small model that is easy to interpret, i.e. in cases where understanding the predictor-response relation is the primary goal.
- **AIC** is used when the principal aim is the prediction of future observations. In these cases, small out-of-sample error is key, but neither the number or meaning of the predictors.

Alternative Search Heuristics

Forward Selection

1) Start with an empty model, i.e. only the intercept, but no predictors. The fitted value is the mean of the responses.
2) In a step-by-step manner, the predictor which leads to the best AIC/BIC value is added to the model.
3) Adding predictor variables is repeated until the AIC/BIC value can no longer be improved.

R: `> fit.aic <- step(fit, dir="forward", k=2)`

→ Forward Selection is used with big datasets, where backward elimination is too time consuming.

Alternative Search Heuristics

Stepwise Model Search

- This is an alternation of forward and backward steps. We can either start with the full model (1. step is backwards) or with the empty model (1. step is forward).
- In each forward step, all predictors can be added, also these that were excluded before. In each backward step, any of the predictors can be kicked out of the model (again).

→ Similar to Backward Elimination resp. Forward Search

→ Not much more time consuming, but more exhaustive

→ Recommended!

Stepwise Model Search in R

Starting with an empty model:

```
> null <- lm(Mortality ~ 1, data=mortality)
> all  <- lm(Mortality ~ ., data=mortality)
> fit  <- step(null, scope=list(upper=all))
```

Starting with the full model:

```
> fit  <- step(all, direction="both", k=2)
```

Note:

Argument `scope=...` allows specifying arbitrary minimal and maximal models for both cases. Then some predictors can be added or be removed from the model.

Alternative Search Heuristics

All Subsets Regression

- When m predictors are present, there are in fact 2^m different models that could be tried for finding the best one.
- In cases where m is small (i.e. $m \approx 10-20$) all submodels (up to a certain size) can be tried and evaluated by computing the AIC/BIC criterion.

→ Complete search, but enormous computing time needed

→ Yields a good solution, but not the causal model either

→ Recommended for small dataset where it is feasible

→ R implementation with function `leaps()`

All Subsets Regression in R

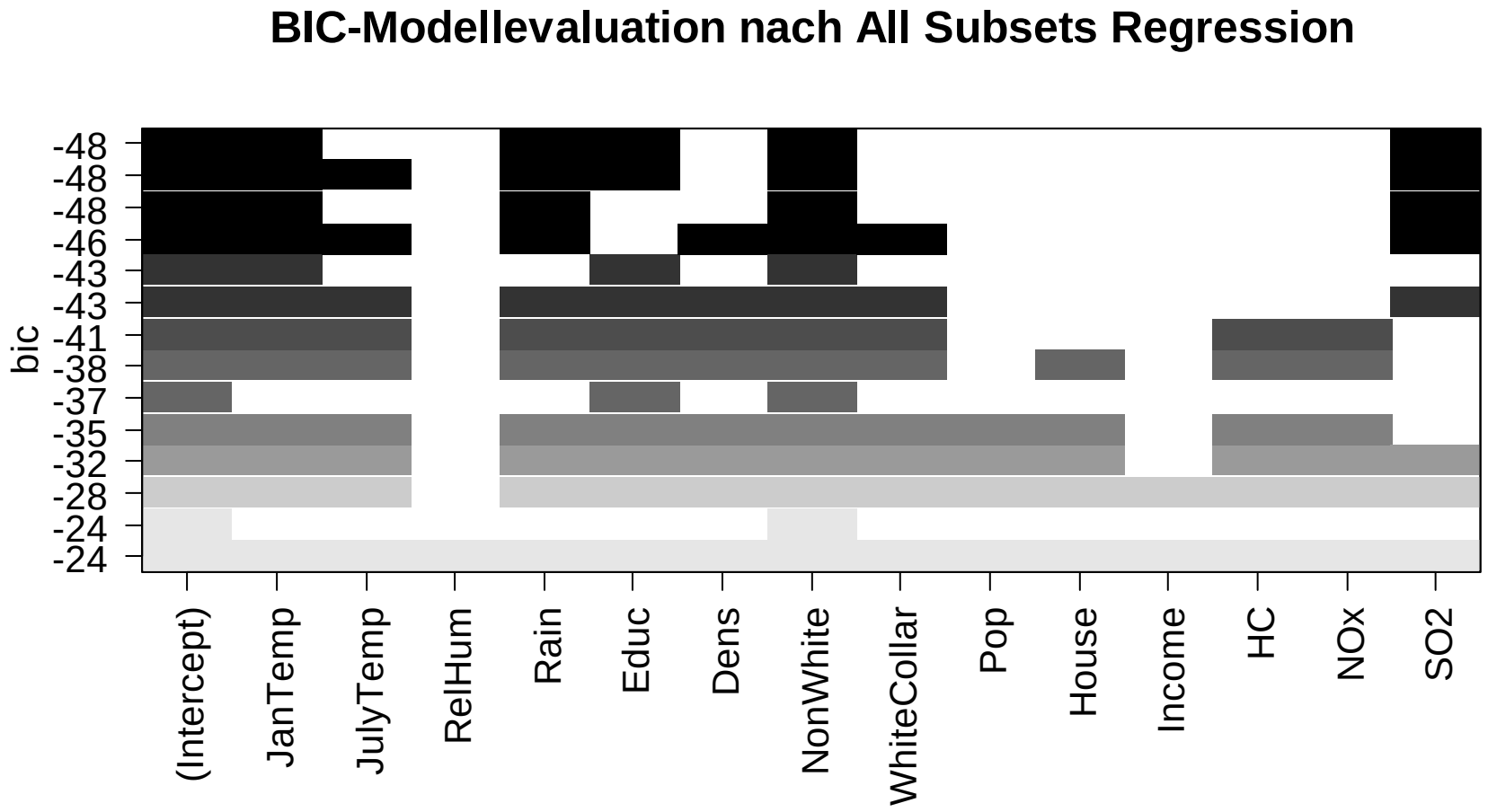
R commands:

```
> library(leaps)
> out <- regsubsets(Mortality~., nbest=1,
                    data=mortality, nvmax=14)
> summary(out)
> plot(out)
```

Note:

The procedure starts with the empty model and for each number of predictors, identifies the `nbest=1` models. By typing `~.` in the formula, all predictors are allowed. The maximum model size that is search can be determined with `nvmax=14`.

Applied Statistical Regression
AS 2014 – Multiple Regression
Visualization of All Subsets Selection



Final Remarks

- Each search heuristics yields a different *"best model"*.
- If we had another data sample from the same population and would repeat the variable selection using the same heuristic, the result might turn out to be different.
- The *"best model"* has the character of a random variable.

How to deal with that in practice?

We should not only consider the one "best model" according to a particular search heuristic, but a number of alternative model with similar performance (if they exist).

Interactions and Categorical Input

Models with Interactions

Do not remove main effect terms if there are interactions with these predictors contained in the model.

Categorical Input

- If a single dummy coefficient is non-significant, we cannot just kick this term out of the model, but we have to test the entire block of indicator variables.
- When we work manually and testing based, this will be done with a partial F-test. When working criterion based, `step()` does the right thing

Cross Validation

Definition:

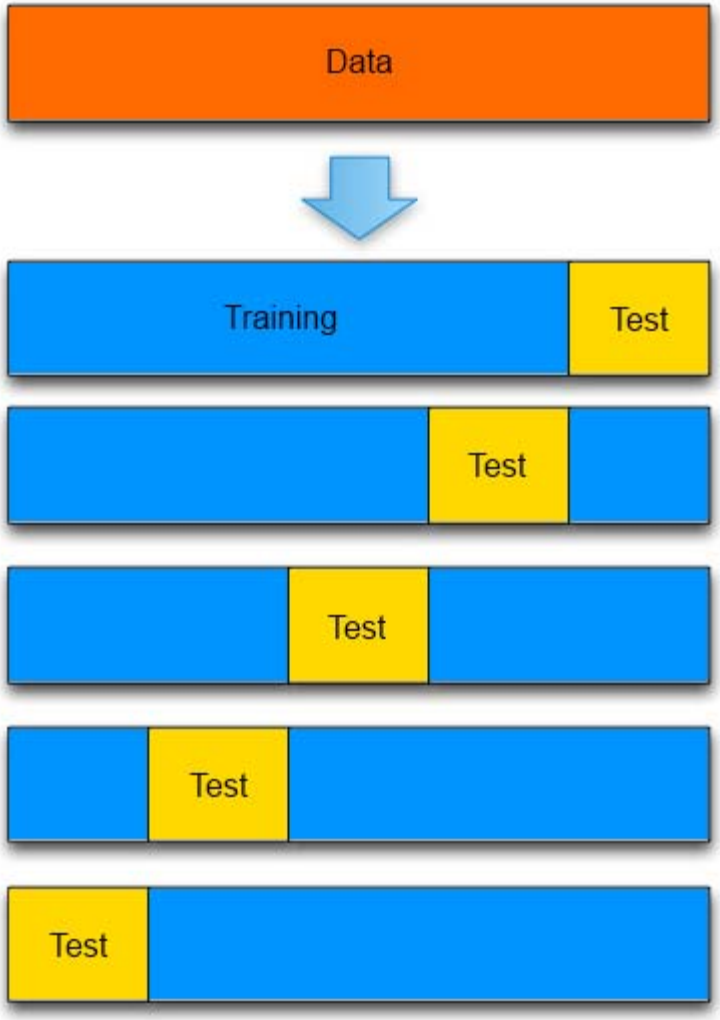
Cross Validation is a *technique for estimating the performance of a predictive model*, i.e. assessing how the prediction error will generalize to a new, independent dataset.

Rationale:

Cross Validation serves to prevent *overfitting*. On a given data set, a *bigger model* always yields a *better fit*, i.e. *smaller RSS*, *higher R-squared*, *less error variance*, et cetera.

While *AIC/BIC* and the *adjusted R-squared* try to overcome this problem by penalizing for model size, its use is limited in reality.

Cross Validation: How It Works…



- In this schematic example, 5-fold CV is illustrated.
- Each observations belongs to exactly 1 test set. The test sets are of roughly equal size.
- Also, each data point belongs to exactly 4 training sets.
- In each fold, the test RSS is recorded. The CV-RSS is the summed result over all folds.

When to Use Cross Validation

- If the *ultimate goal* is *predicting new data points*, then it is self suggesting to identify a model which does well at this. We can mimic the prediction task on our training sample with *cross validation.*

AIC/BIC and the adjusted R-squared do not work in case of:

- *Response variable transformation:* if one wants to find out whether a model with logged or non-logged response is better for prediction, cross validation is the only option.
- *Non-identical data:* if we want to evaluate the quality of a fitted with or without removing sketchy data points, then also cross validation is our only option.

Cross Validation

Further remarks:

- Cross validation evaluates the out-of-sample performance, i.e. how precisely a model can forecast observations that were not used for fitting the model.
- There are alternatives to 5-fold CV. Popular is n-fold CV, which is known as *Leave-One-Out Cross Validation*.
- In R, it's easy to code "for-loops" that do the job, but there are also existing functions (that have some limits...):

```
> library(DAAG)
> CVlm(data, formula, fold.number, …)
```

Cross Validation

Using `for()` to program cross validation loops:

```
> rss   <- c()
> fo    <- 5
> sb    <- round(seq(0,nrow(dat),length=(fo+1)))
> for (i in 1:folds)
> {
>   test   <- (sb[((fo+1)-i)]+1):(sb[((fo+2)-i)])
>   train  <- (1:nrow(dat))[-test]
>   fit    <- lm(res ~ p1+..., data=dat[train,])
>   pred   <- predict(fit, newdata=dat[test,])
>   rss[i] <- sum((dat$response[test] - pred)^2)
> }
```

Applied Statistical Regression
AS 2014 – Multiple Regression

Modelling Strategies

We have learnt a number of technical details about multiple linear regression. The often asked question is in which order the tools need to be applied:

Transformation → Estimation → Model Diagnostics → Variable Refinement & Selection → Evaluation

This is a good generic solution, but not an always-optimal strategy.

Professional regression analysis is the search for structure in the data. It requires technical skill, flexibility and intuition. The analyst must be alert to the obvious as well as to the non-obvious, and needs the flair to find the unexpected.

Applied Statistical Regression
AS 2014 – Multiple Regression

Modelling Strategies

0) Data Screening & Processing

- learn the meaning of all variables
- give short and informative names
- check for impossible values, errors
- if they exist: set them to NA
- systematic or random missings?

1) Transformations

- bring all variables to a suitable scale
- use statistical and specific knowledge
- routinely apply the log-transformation
- break obvious collinearity

Applied Statistical Regression
AS 2014 – Multiple Regression

Modelling Strategies

2) Fitting a Big Model

Fit a big model with potentially too many predictors

- use all if $p < n/5$!!!
- *or* preselect manually according to previous knowledge
- *or* preselect with forward search and a p-value of 0.2

3) Model Diagnostics

- generate the 4 standard plots in R
- a systematic error is non-tolerable, improve the model!!!
- be aware to influential data points, try to understand them
- take care with non-constant variance & long-tailed errors
- think about potential correlation in the residuals

Applied Statistical Regression
AS 2014 – Multiple Regression

Modelling Strategies

4) Variable Selection

- try to reduce the model to what is utterly required
- run a stepwise search from the full model with AIC/BIC
- if feasible, an all-subset-search with AIC/BIC is even better
- the residual plots must not (substantially) degrade in quality!

5) Refining the Model

- use partial residual plots or plots against other variables
- think about potential non-linearities/factorization in predictors
- interaction terms can improve the fit drastically
- are there still any collinearities that disturb?

Applied Statistical Regression
AS 2014 – Multiple Regression

Modelling Strategies

6) Plausibility

- implausible predictors, wrong signs, results against theory?
- remove if (appropriate) and no drastic change in outcome

7) Evaluation

- cross validation for model comparison & performance
- derive test results, confidence and prediction intervals

8) Reporting

- be honest and openly report manipulations & decisions
- regression models are descriptive, but not causal!
- do not confuse significance and relevance!

Significance vs. Relevance

The larger a sample, the smaller the p-values for the very same predictor effect. Thus do not confuse small p-values with an important predictor effect!!!

With large datasets, we can have:

- statistically significant results which are practically useless
- e.g. high evidence that the response value is lowered by 0.1% which is often a practically totally meaningless result.

Bear in mind that generally:

- most predictors have influence, thus $\beta_j = 0$ hardly ever holds
- the point null hypothesis is thus usually wrong in practice
- we just need enough data so that we are able to reject it

Significance vs. Relevance

Absence of Evidence $\neq$ Evidence of Absence

- if one fails to reject a null hypothesis $\beta_j = 0$ we do not have a proof that the predictor does not influence the response.

- things may change if we have more data, or even if the data remain the same, but the set of predictors is altered.

Measuring the Relevance of Predictors:

- maximum effect of a predictor variable on the response: $\beta_j \cdot (\max_i x_{ij} - \min_i x_{ij})$
- this can be compared to the total span in the response, or it can be plotted vs. the (logarithmic) p-value.

Applied Statistical Regression
AS 2014 – Multiple Regression

Mortality: Which Predictors Are Relevant?

```
> summary(fit.step)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1031.9491    80.2930  12.852  < 2e-16 ***
JanTemp       -2.0235     0.5145  -3.933  0.00025 ***
Rain           1.8117     0.5305   3.415  0.00125 **
Educ         -10.7463     7.0797  -1.518  0.13510
NonWhite       4.0401     0.6216   6.500  3.1e-08 ***
WhiteCollar   -1.4514     1.0451  -1.389  0.17082
log(Nox)      19.2481     4.5220   4.257  8.7e-05 ***
---
Residual standard error: 33.72 on 52 degrees of freedom
Multiple R-squared: 0.7383,  Adjusted R-squared: 0.7081
F-statistic: 24.45 on 6 and 52 DF,  p-value: 1.543e-13
```

Mortality: Which Predictors Are Relevant?

Implementing the idea of maximum predictor effect:

```
> mami   <- function(col) max(col)-min(col)
> ranges <- apply(mort,2,mami)[c(2,5,6,8,9,14)]
> ranges
JanTemp    Rain  Educ  NonWhite  WhiteCollar  log.NOx
  55.00   55.00  3.30     37.70        28.40     5.77
>
> rele   <- abs(ranges*coef(fit.step)[-1])
> rele
JanTemp    Rain  Educ  NonWhite  WhiteCollar  log.NOx
 111.29   99.64 35.46    152.31        41.22   110.97
```

Predictor contributions are quite evenly distributed here.
Maximum span in the response is **322.43**

What is a Good Model?

- The *true model* is a concept that exists in theory & simulation, but whether it does in practice remains unclear. Anyway, it is not realistic to identify the true model in observational studies.
- A **good model** is *useful* for the task at hand, *correct*ly describes the data without any systematical errors, has good *predictive* power and is *practical*/applicable for future use.
- Regression models in observational studies are *always only descriptive, but never causal*. A good model yields an accurate idea which of the observed variables drives the variation in the response, but not necessarily reveals the true mechanisms.