5 Model development

5.1 The Task

a Which explanatory variables shall appear in the model formula in which form?

b Example Construction Cost of nuklear power plants

	Explanation	Type	Trsf.
K	Construction Cost	amount	log
G	Capacity	amount	log
D	Date of permission	contin.	–
WZ	Waiting time betw. application & perm.	amount	–
BZ	Construction time	amount	–
Z	Follow-up plant (existing plant on site)	binary	–
NE	Site in NE of the US	binary	–
KT	Cooling tower	binary	–
BW	Reactor by Babcock-Wilcox	binary	–
N	Number of plants built by the same architects/engineers earlier, +1	count	sqrt
KG	Partial price guarantee of the project leading enterprise	binary	–

5.1

c First aid transformations:

d $log_{10}\langle \mathtt{K}\rangle = \beta_0 + \beta_1\, log_{10}\langle \mathtt{G}\rangle + \beta_2 \mathtt{D} + \beta_3 \mathtt{WZ} + \beta_4 \mathtt{BZ}$
$+\beta_5 \mathtt{Z} + \beta_6 \mathtt{NE} + \beta_7 \mathtt{K} + \beta_8 \mathtt{BW} + \beta_9 \sqrt{\mathtt{N}} + \beta_{10} \mathtt{KG} +$ Fehler

e A single term

- t test for a β_j

- factor (nominal variable)

  $\longrightarrow$ F test for several β_j

  Does a significance test make sense in this context?

Coefficients:

	Value	Std. Error	t value	Pr($> \|t\|$)	Signif
(Intercept)	-6.02586	2.34729	-2.57	0.018	*
lg10(G)	0.69254	0.13713	5.05	0.000	***
D	0.09525	0.03580	2.66	0.015	*
WZ	0.00263	0.00955	0.28	0.785	
BZ	0.00229	0.00198	1.16	0.261	
Z	-0.04573	0.03561	-1.28	0.213	
NE	0.11045	0.03391	3.26	0.004	**
KT	0.05340	0.02970	1.80	0.087	.
BW	0.01278	0.04537	0.28	0.781	
sqrt(N)	-0.02997	0.01780	-1.68	0.107	
KG	-0.09951	0.05562	-1.79	0.088	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

5.2 Automatic Model Selection

a Stepwise backwards

b delete `WZ` !

delete `BW`, `BZ`, `Z`, $\sqrt{\mathtt{N}}$ und `KT` !

Coefficients:

	Value	Std. Error	t value	$\Pr(> \|t\|)$	Signif
(Intercept)	-3.4612	1.1458	-3.02	0.005	**
log10(G)	0.6629	0.1295	5.12	0.000	***
D	0.0610	0.0160	3.82	0.001	***
NE	0.0831	0.0330	2.52	0.018	*
KG	-0.1844	0.0424	-4.35	0.000	***

c Stepwise forward ...

e **all subsets**.

f **Criteria**

1. „Coefficient of Determination" R^2 or multiple correlation R,
2. Value of F test statistic for the model,
3. P value for F test,
4. Estimated variance of error $\widehat{\sigma}^2$,

g

5. “Adjusted” coef. of det.: $R^2_{\text{adj}} = 1 - \frac{n-1}{n-p'}(1 - R^2)$
6. $C_p := \text{SSQ}^{(E)}/\widehat{\sigma}_m^2 + 2p' - n = n(\text{MS}_E/\widehat{\sigma}_m^2 - 1 + 2p'/n)$,
7. Akaike’s Information criterion $\text{AIC} \approx C_p$.

Larger model are not always better!

5.2

h C_p in the example:

Add $\mathtt{KT}$ and $\sqrt{\mathtt{N}}$!

P value for KG: 0.049.

i Lasso Penalized Regression: Penalty on large coefficients

Minimize

$$Q\left\langle \underline{\beta};\lambda\right\rangle = \sum\nolimits_i R_i^2 + \lambda \sum\nolimits_j \left|\beta_j\right| .$$

λ: weight of penalty

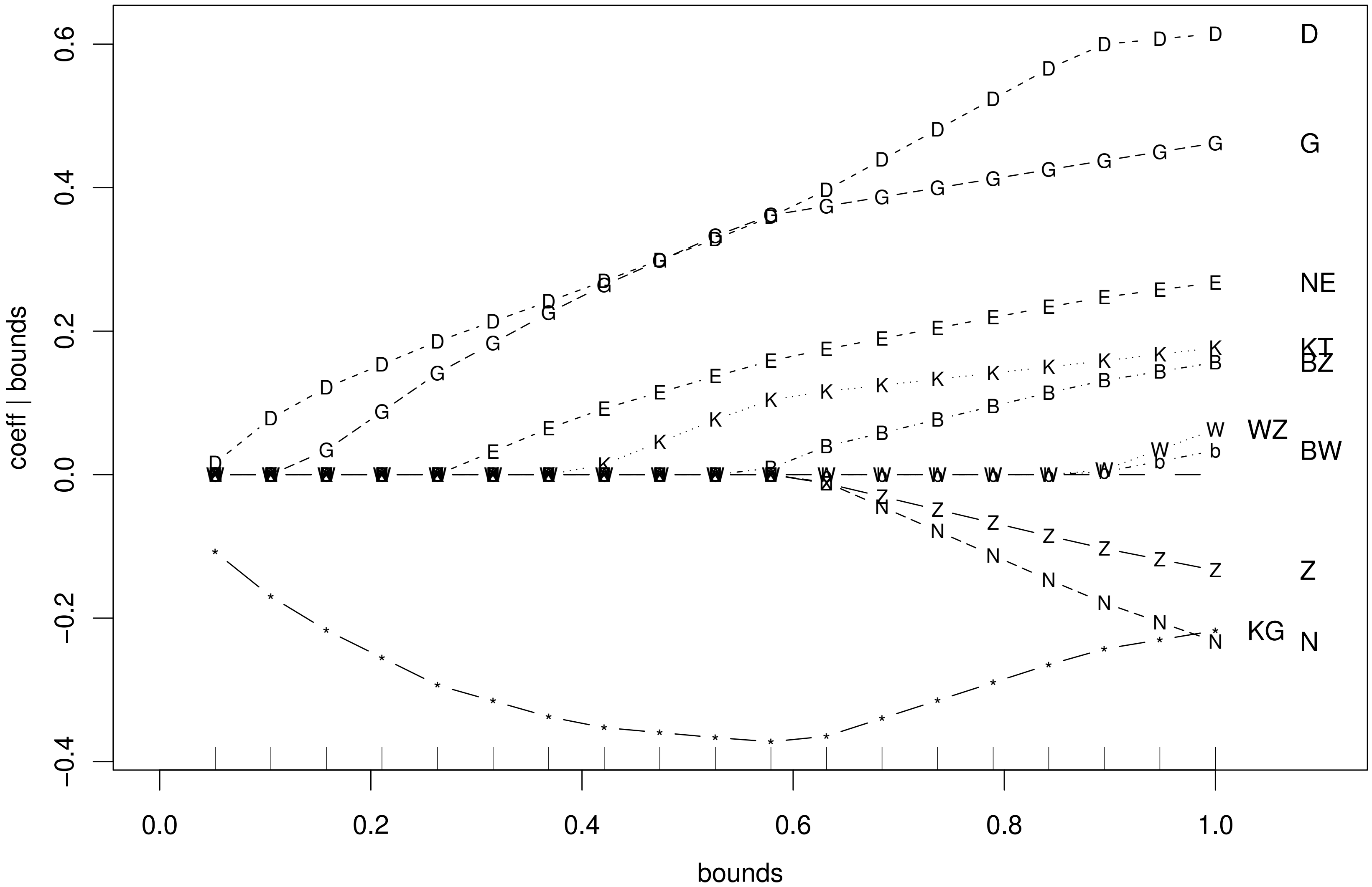
Variation of λ $\longrightarrow$ some coefficients =0

$\longrightarrow$ Model selection

- Standardize $X^{(j)}$ to render β_j comparable.
- Adaptive Lasso:

  Attach weights to β_j: weights $1/\widehat{\beta}_j$ with
  $\widehat{\beta}_j$ from preliminary Lasso estimate.
- Lasso is suitable for large numbers of $X^{(j)}$, $p >> n$.

  $\longrightarrow$ Genomics, Proteomics


coeff | bounds
bounds
0.6
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
D
G
NE
KT
BZ
WZ
BW
Z
KG
N

j Choice of the weight of the L1 penalty term, λ

“Cross validation”, 10-fold.

k Is the “best” model the true model?

Consider several models as the result of the analysis

Among all “good” models choose one or more suitable one(s)

by plausibility and subject matter knowledge!

Exploratory data analysis will NOT find the “true” model

but several which fit the data well.

5.3 Collinearity

a Model $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{E}$

$\boldsymbol{X}$ is singular, $X^{(j)}$'s collinear if

$$\begin{aligned} \boldsymbol{X} \text{ singulär} \quad &\Longleftrightarrow \quad \det\langle \boldsymbol{X} \rangle = 0 \\ \Longleftrightarrow \quad &\text{es gibt } \underline{c} \text{ mit } \boldsymbol{X}\underline{c} = \underline{0} \quad (\underline{c} \neq \underline{0}) \\ \Longleftrightarrow \quad &\text{es gibt ein } j \text{ mit } \quad x_i^{(j)} = \widetilde{c}_0 + \sum_{k \neq j} \widetilde{c}_k x_i^{(k)} \end{aligned}$$

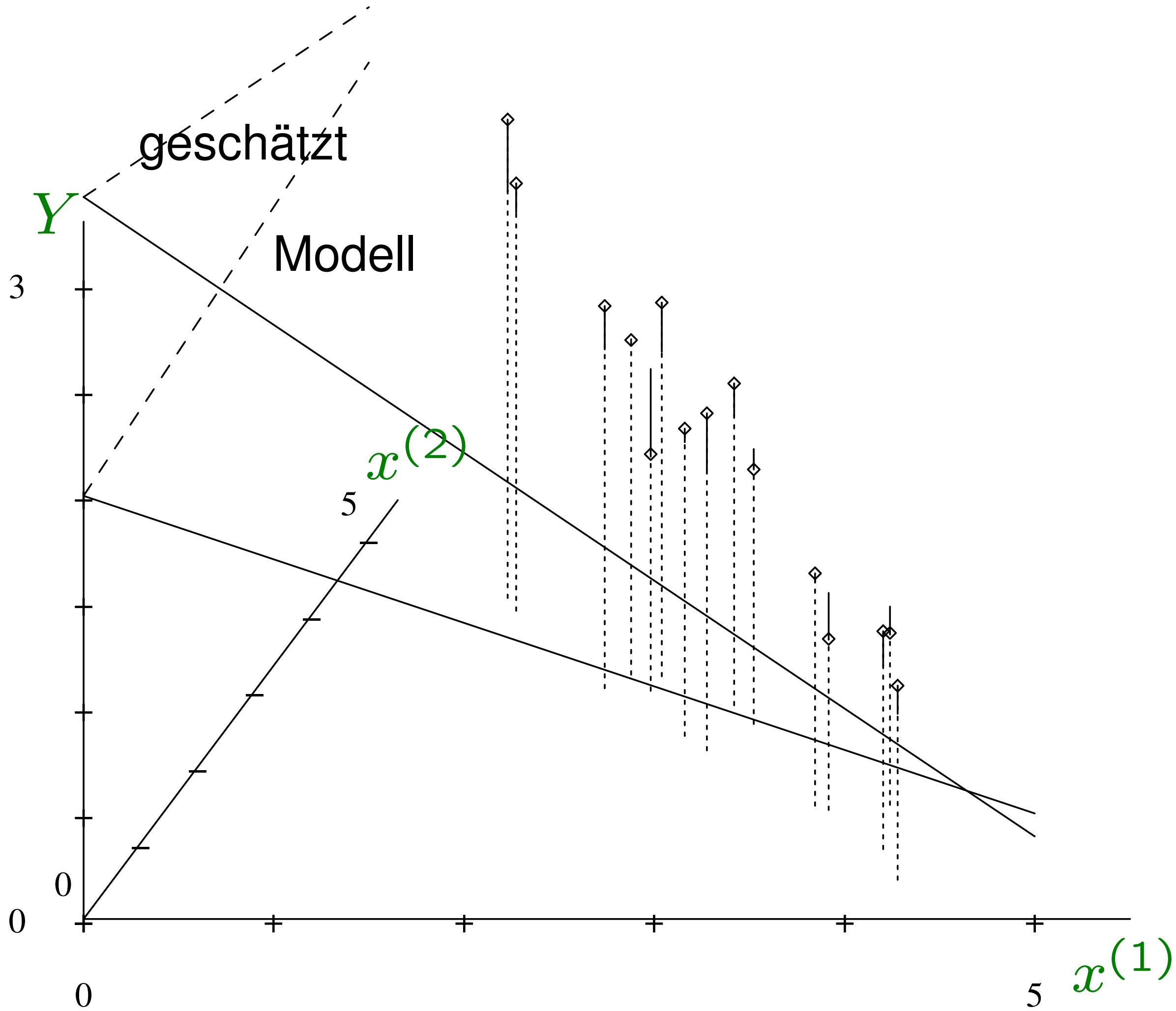
Parameter not unique since

$$\boldsymbol{X}\underline{\beta} = \boldsymbol{X}(\underline{\beta} + \gamma\underline{c}), \quad \gamma \text{ beliebig}$$

b Solution: Delete a column!

Caution: Interpretation of parameters may change!

c Approximate collinearity $\longrightarrow$ parameter ill determined



d Large standard errors of estimates $\longrightarrow$ coefficients insignificant

e How to detect collinearity?

– Standard error of the $\widehat{\beta}_j$'s

– Is there a relation $x_i^{(j)} \approx \widetilde{c}_0 + \sum_{k \neq j} \widetilde{c}_k x_i^{(k)}$?

= Regression problem! Coefficient of determination R_j^2

or variance inflation factor $\mathrm{VIF}_j = 1/(1 - R_j^2)$

5.3

f **What remedies against collinearity?**

– Choice of experimental conditions,

g – linear transformation of x sup j's, e.g., sum and difference

or “more important” variable plus residuals of the other one.

h – delete variable wiht highest R_j^2! (Usually insignifikant!)

i* **Ridge Regression** = Penalized Regression. – Penalize squared β_j:

$$Q\left\langle \underline{\beta}; \lambda \right\rangle = \sum\nolimits_i R_i^2 + \lambda \sum\nolimits_j \beta_j^2 \ .$$

5.4 Strategies of Model Development

a Model selection is an interplay between

- available knowledge from subject matter & statistics,
- Residual analysis, „detektive work",
- automatic model selection procedures,
- Residual analysis, „detektive work",
- Prinziple of simplicity,
- Assessment of plausibility and
  critique by subject matter knowledge.

0. Read data, define variable names (sounds trivial...)

check plausibility (screening), get acquainted

1. “First aid” Transformations.

2. A large model

- all variables (main effects),
- Result of a stepwise forward selection

3. Examination of the Random part:

- Outliers in residuals,
- Distribution of residuals,
- Equality of variances,
- Independence of errors.

It may be warranted in view of the results to

- transform the target variable,
- introduce weights,
- use robust methods (if not done routinely)

4. Non-linearities.

5. Automatic model selection

6. Add variables

7. Interactions

8. Influential observations

9. Critique by subject matter knowledge

10. Examine fit

11. Revision

12. Check deleted terms again

Celebrate!

5.4

b Example construction cost

Question: Does price guarantee help?

Detective word gives the most convincing answer!

Messages

1. Automatic model selection procedures are a helpful tool
   but do not find “the truth!”

2. Model selection is an interplay between

   - available knowledge from subject matter & statistics,
   - Residual analysis, „detektive work",
   - automatic model selection procedures,
   - Residual analysis, „detektive work",
   - Principle of simplicity,
   - Plausibility & critique by subject matter knowledge.