

## 5 Modell-Entwicklung

### 5.1 Problemstellung

- a Welche Ausgangs-Variablen sollen in welcher Form in der Modell-Gleichung der linearen Regression erscheinen?
- b **Beispiel Baukosten**

| Bez. | Bedeutung   | Typ     | Transf. |
|------|---|---------|---------|
| K    | Baukosten   | Betrag  | log     |
| G    | Grösse  | Betrag  | log     |
| D    | Datum der Baubewilligung  | kontin. | —       |
| WZ   | Wartezeit zwischen Antrag und Baubewilligung                        | Betrag  | —       |
| BZ   | Bauzeit: Zeit bis Inbetriebnahme                                    | Betrag  | —       |
| Z    | Zweitwerk: früheres Werk auf gleichem Gelände                       | binär   | —       |
| NE   | Werk steht im Nordosten der USA                                     | binär   | —       |
| KT   | Werk arbeitet mit Kühlturm  | binär   | —       |
| BW   | Reaktor hergestellt durch Babcock-Wilcox                            | binär   | —       |
| N    | Anzahl Werke, die das gleiche Ingenieur-Team bereits erbaut hat, +1 | Anzahl  | Wurzel  |
| KG   | Partielle Kostengarantie des Generalunternehmers                    | binär   | —       |

## 5.1

c First aid transformations:

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

e **Ein einzelner Term.**

- t-Test für einzelnes  $\beta_j$
- F-Test für mehrere  $\beta_j$  aufs Mal,  
z. B. alle  $\beta_j$  für eine nominale Variable.

F-Test für den Vergleich von Modellen:

## 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 Automatisierte Verfahren zur Modellwahl

a **Schrittweise rückwärts**

b WZ weglassen!

BW , BZ , Z ,  $\sqrt{N}$  und KT weglassen!

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 **Schrittweise vorwärts** ...

e „Alle Gleichungen“ – all subsets.

f Kriterien

1. „Bestimmtheitsmass“  $R^2$  oder multiple Korrelation  $R$ ,
2. Wert der Test-Statistik für das gesamte Modell (F-Test),
3. zur F-Test-Statistik gehöriger P-Wert,
4. geschätzte Varianz  $\hat{\sigma}^2$  der Fehler
- 9 5. Korrigiertes Bestimmtheitsmass  $R^2$  (adjusted  $R^2$ ):
 
$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-p'}(1 - R^2)$$
6.  $C_p := SSQ^{(E)} / \hat{\sigma}_m^2 + 2p' - n = n(\text{MSE} / \hat{\sigma}_m^2 - 1 + 2p'/n)$ ,
7. Informations-Kriterium von Akaike  $\text{AIC} \approx C_p$ .

Grössere Modelle sind nicht immer besser.

## 5.2

h  $C_p$  im Beispiel:

KT und  $\sqrt{N}$  dazunehmen!

P-Wert für KG: 0.049.

- i **Lasso** Penalized Regression: Bestrafe grosse Koeffizienten  
Minimiere

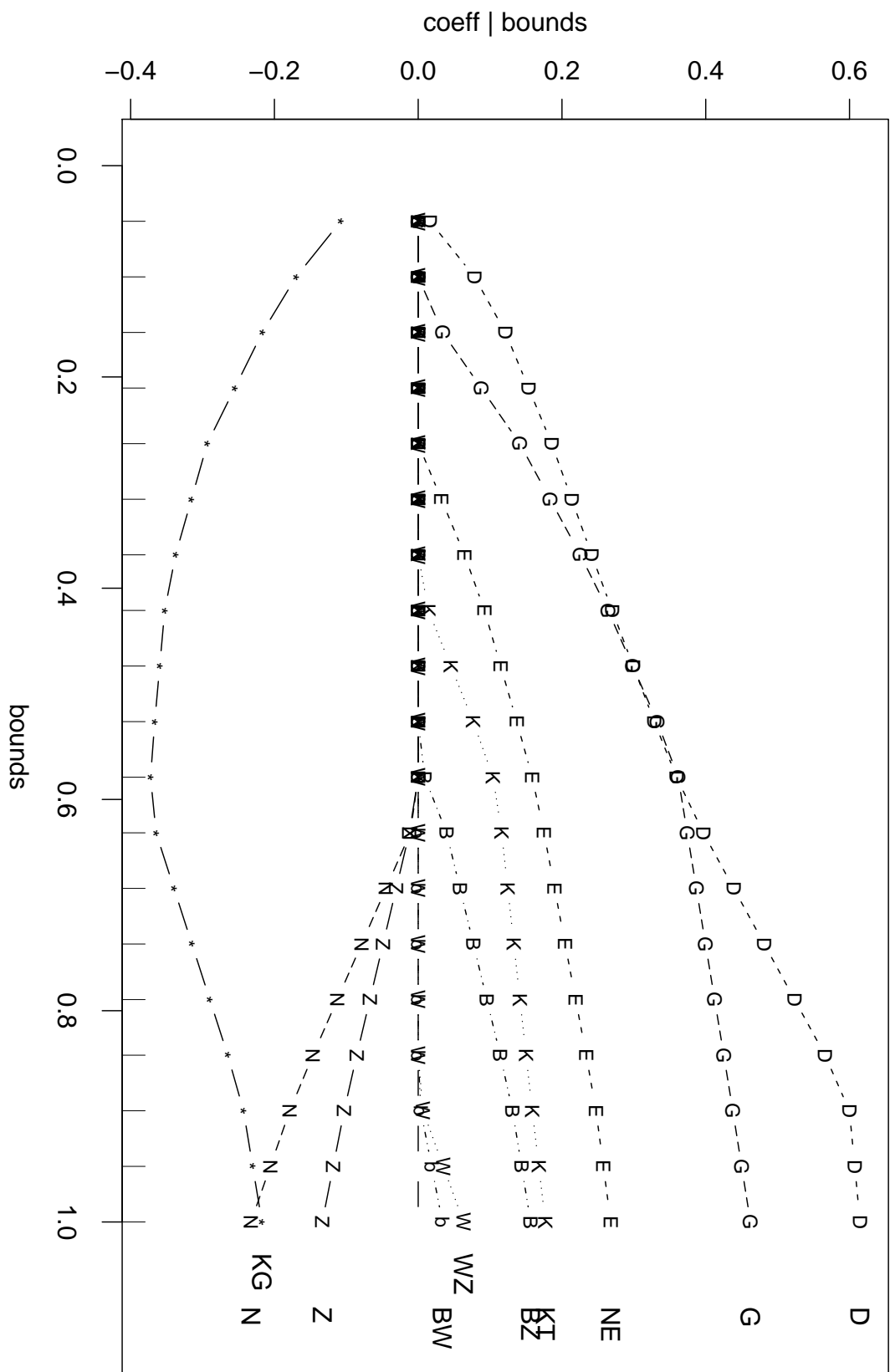
$$Q(\underline{\beta}; \lambda) = \sum_i R_i^2 + \lambda \sum_j |\beta_j|.$$

$\lambda$ : wie stark bestrafen?

Variation von  $\lambda$   $\longrightarrow$  Koeffizienten werden exakt 0.

$\longrightarrow$  Modellwahl

- $X^{(j)}$  standardisieren, um  $\beta_j$  vergleichbar zu machen.
- **Adaptives Lasso:**  
Gewichte  $\beta_j$  mit  $1/\hat{\beta}_j$ ,  $\hat{\beta}_j$  aus erstem Durchgang!
- Lasso bewährt sich für **grosse Anzahl**  $X^{(j)}$ ,  $p \gg n$ .  
 $\longrightarrow$  **Genomics, Proteomics**



## 5.2

- j **Wahl der Gewichts der L1-Bestrafung,**  
“Cross validation”, 10-fold.

## 5.2

k Bestes = wahres Modell ?

Mehrere als Ergebnis des Verfahrens deklarieren!

Unter den „guten“ Modellen soll mit Hilfe von

**Plausibilitäts-Überlegungen und Fachwissen**

ein geeignetes (oder wenige geeignete)

ausgewählt werden.

**Explorative Datenanalyse findet NICHT**

**das richtige Modell, sondern**

einige Modelle, die den Daten gut entsprechen.

## 5.2

- | **Prioritäten von Termen** Prinzip: Wenn quadratischer Term im Modell ist, lässt man den linearen nicht weg.

Wieso? ...

... ausser man habe gute Gründe, vom Prinzip abzuweichen.

Ebenso:

- Wechselwirkung  $X_1 : X_2$  → beide Haupteffekte drin lassen.
- Intercept immer drin behalten.

## 5.3 Kollinearität

a Modell  $\underline{Y} = \mathbf{X} \underline{\beta} + \underline{E}$

$\mathbf{X}$  ist **singulär**,  $X^{(j)}$ 's kollinear, wenn

$$\mathbf{X} \text{ singulär} \iff \det\langle \mathbf{X} \rangle = 0$$

$$\iff \text{es gibt } \underline{c} \text{ mit } \mathbf{X} \underline{c} = \underline{0} \quad (\underline{c} \neq \underline{0})$$

$$\iff \text{es gibt ein } j \text{ mit } x_i^{(j)} = \tilde{c}_0 + \sum_{k \neq j} \tilde{c}_k x_i^{(k)}$$

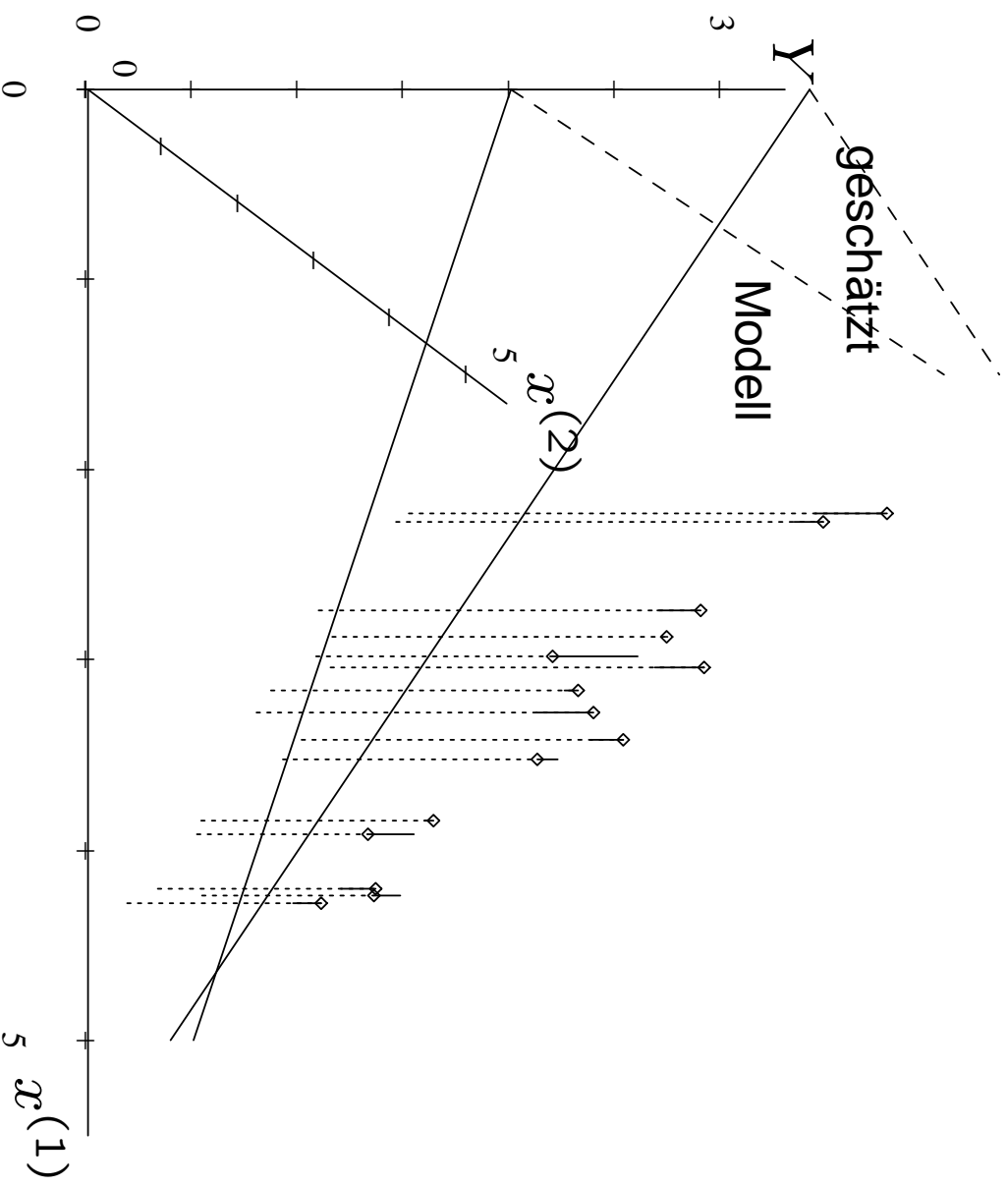
Parameter nicht eindeutig, da

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

b Lösung: Kolonne streichen!

Achtung: Interpretation der Parameter kann sich ändern!

- c Genäherte Kollinearität  $\longrightarrow$  Parameter schlecht bestimmt.



## 5.3

- d Grosse Standardfehler für Schätzungen  $\longrightarrow$  Koeffizient nicht signif.  $\neq 0$ .
- e Vorhersage im Bereich der Daten i.O.
- f **Wie entdeckt man Kollinearität?**
  - Standardfehler der  $\hat{\beta}_j$
  - Gibt es eine Beziehung  $x_i^{(j)} \approx \tilde{c}_0 + \sum_{k \neq j} \tilde{c}_k x_i^{(k)}$  ?  
 = Regressionsproblem! Bestimmtheitsmass  $R_j^2$   
 oder variance inflation factor  $VIF_j = 1/(1 - R_j^2)$

5.3

g **Was tun gegen Kollinearität?**

– Wahl der Versuchsbedingungen.

h –  $x$ -Variable linear transformieren, z.B. Summe und Differenz oder „wichtigere“ Variable und Residuen der anderen.

i – Variable mit dem höchsten  $R_j^2$  wegl! (Meist sowieso nicht signifikant)

j\* **Ridge Regression** = Penalized Regression. – Bestrafung mit quadr.  $\beta_j$ :

$$Q(\underline{\beta}; \lambda) = \sum_i R_i^2 + \lambda \sum_j \beta_j^2.$$

## 5.4 Strategien der Modell-Entwicklung

- a Modellwahl ist ein Zusammenspiel von
- Vorwissen aus Anwendung und Statistik,
  - Residuen-Analyse, „Detektivarbeit“,
  - automatischen Modellwahl-Methoden,
  - Residuen-Analyse, „Detektivarbeit“,
  - Prinzip der Einfachheit,
  - **Beurteilung der Plausibilität vom Fachwissen her.**

**Modellwahl ist von der Fragestellung abhängig!**

- (a) Welche Variablen beeinflussen die Zielgrösse?  
→ Achtung! Es gibt nur Indizien!
- (b) **Vorhersage.**
- (c) **Modell im Wesentlichen vorgegeben.**  
Allenfalls **Störvariable** einbeziehen.  
Beispiel Medikamentenprüfung.

Im Folgenden soll (b) oder (a) gefragt sein.

0. Daten einlesen, Variablennamen festlegen  
„plausibilisieren“ (screening), kennenlernen
1. „First aid“ Transformationen.
2. Ein grosses Modell
  - alle Variablen (Haupteffekte),
  - Ergebnis eines „Schrittweise-Vorwärts-Verfahrens“

### 3. Überprüfung des zufälligen Teils:

- Ausreisser in den Residuen,
- Verteilung der Residuen,
- Gleichheit der Varianzen,
- Unabhängigkeit der Fehler.

Es kann aufgrund der Ergebnisse angezeigt sein,

- die Zielgrösse zu transformieren,

- Gewichte einzuführen,
  - robuste(re) Methoden zu verwenden, soweit dies nicht schon sowieso geschieht.
4. **Nicht-Linearitäten.** Residuen gegen Ausgangsgrößen.
  5. **Automatisierte Variablen-Wahl**
  6. **Variable hinzufügen.**
  7. **Wechselwirkungen.** Erst nach Bereinigung der Nicht-Lin. durch entspr. Plots oder numerisch  
`step ( . . . , scope = ( x1+x2+ . . . ) ^ 2 )`

8. Einflussreiche Beobachtungen.
9. Kritik mit Fachwissen.
10. Anpassung prüfen.
11. Revision.
12. Entfernte Terme überprüfen.

Feiern!

5.4

b **Beispiel der Baukosten**

Frage nach dem Nutzen der Kostengarantie

Hier führt Detektivarbeit zur überzeugendsten Antwort!

## Merkpunkte

## Modell-Entwicklung

1. **Automatisierte** Verfahren zur Variablenwahl sind ein nützliches Hilfsmittel – finden aber nicht „die Wahrheit“
2. Modellwahl ist ein **Zusammenspiel** von
  - Vorwissen aus Anwendung und Statistik,
  - **Residuen-Analyse, „Detektivarbeit“**,
  - automatischen Modellwahl-Methoden,
  - Residuen-Analyse, „Detektivarbeit“,
  - Prinzip der Einfachheit,
  - **Beurteilung der Plausibilität vom Fachwissen her.**