



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

Institut für Statistik



# Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2014



- 5 Metrische Einflußgrößen: Polynomiale Regression, Trigonometrische Polynome, Regressionsplines, Transformationen.
- 6 Modelldiagnose
- **Variablenselektion**
- 8 Das allgemeine lineare Modell: Gewichtete KQ-Methode, Autokorrelierte und heteroskedastische Störterme
- 9 Das logistische Regressionsmodell
- 10 Das gemischte lineare Regressionsmodell („Linear mixed Model“)

# Modellwahl: Zielsetzung der Modellierung

---

- a) gute Beschreibung des Verhaltens der Zielgröße
- b) Vorhersage zukünftiger Werte der Zielgröße und Schätzung des Mittels der Zielgröße
- c) Extrapolation auf Bereiche außerhalb der X-Daten
- d) Schätzung von Parametern
- e) Kontrolle eines Prozesses durch Variation des Inputs  
→ Kausalität nötig
- f) Entwicklung realistischer Modelle für einen Prozess  
→ Kausalität nötig

# Grundsätzliches I

---

Allgemein: „Tradeoff“ zwischen Modell-Genauigkeit ( $R^2$ ) und Einfachheit.

Viele Variablen →  $R^2$  steigt  
→ Komplexität nimmt zu

a) b) Modellgenauigkeit wichtig

→ viele Variablen

→ Kausale Beziehung nicht nötig „Variable enthält Information“

c) Extrapolation

→ Kausalität

# Grundsätzliches II

---

d) Bias durch Weglassen von Variablen

Erhöhung der Varianz von Schätzern durch überflüssige Variablen

Beachte: Interpretation der Regressionskoeff. „bei Festhalten der anderen Variablen“.

→ Einschränkung durch viele Kovariablen.

e) Kontrolle

→ Kausalität erforderlich

Realistische Beschreibung

→ Sparsames klares Modell

# Grundsätzliches III

---

Kein Verfahren kann (zunächst bei den Zielen (c, d, e, f))  
das Fachwissen ersetzen

→ Verfahren eher explorativ.

Bei der Prognose können bei größeren Datenmengen  
Variablenselektionsverfahren sehr effizient sein.

# Weitere Aspekte

---

- 1 Diskrete Variablen bereiten zusätzliche Probleme:  
Gleiches gilt für Interaktionseffekte  
Es gibt „Regeln“ z.B.,
  - a) Interaktionen nicht ohne Haupteffekte ins Modell
  - b) Effekte von kategoriellen Variablen nur als Ganzes ins Modell

Aber andere Möglichkeit: Verwenden von Indikator-Variablen

→ Indikator-Variablen, die nicht im Modell sind, entsprechen der gemeinsamen Referenzkategorie

- 2 Multikollinearität kann ein erhebliches Problem sein.  
→ Korrelation der Kandidaten - Einflussgrößen analysieren
- 3 Transformation, quadratische Terme liefern weitere Möglichkeiten

# Maße für die Modellgüte I

---

Gegeben sei das lineare Modell  $Y = X\beta + \varepsilon$

**a) Bestimmtheitsmaß:**

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST} \quad (7.1)$$

**b) Adjustiertes Bestimmtheitsmaß:**

$$R_{adj}^2 := 1 - \frac{MSE}{MST} = 1 - \frac{\hat{\sigma}^2}{SST/(n-1)} \quad (7.2)$$



# Maße für die Modellgüte II

---

## c) Akaikes Informationskriterium

$$AIC = n \ln(SSE) + 2p' - n \ln(n) \quad (7.3)$$

Ausgleich zwischen Anpassung (SSE) und Komplexität ( $2p'$ )

## d) Schwarz'sches Bayes-Kriterium SBC (=BIC)

$$SBC = n \ln(SSE) + \ln(n)p' - n \ln(n) \quad (7.4)$$

Komplexität wird stärker gewichtet als bei AIC

# Maße für die Modellgüte III

---

e) Mallows  $C_p$

$$C_p = \frac{\text{SSE}}{\hat{\sigma}_G^2} + 2p' - n \quad (7.5)$$

$\hat{\sigma}_G$ : Schätzung aus vollem Modell

# Variablenselektionsverfahren I

---

Gegeben ist eine Zielgröße  $Y$  und mehrere mögliche Einflussgrößen  $x_k, k = 1, \dots, K$ . Gesucht ist ein möglichst gutes Modell

$$Y = \beta_0 + \sum_{j=1}^L \beta_j x_{k_l},$$

wobei die  $x_{k_l}, l = 1, \dots, L$  ausgewählt werden sollen.

## 1 Auswahl nach einem Kriterium

Wähle aus allen möglichen  $2^k$  Modellen das Modell mit optimalem Kriterium  $C$  aus.  $C$  ist in der Regel ein Kriterium aus (7.1) - (7.4).

## 2 Vorwärtsselektion

- a) Wähle im Anfangsschritt das Modell  $Y = \beta_0$ .
- b) Im ersten Schritt wird die Variable in das Modell aufgenommen, die zu dem höchsten  $R^2$  führt.
- c) In den weiteren Schritten wird jeweils eine Einflussgröße in das Modell zusätzlich aufgenommen. Es wird jeweils die Variable, die zu dem höchsten  $R^2$  des resultierenden Modells führt, aufgenommen.
- d) **Stoppregel:** Die Prozedur wird beendet, falls ein bestimmtes Zielkriterium erfüllt ist, z.B. AIC

## 3 Rückwärts-Selektion

- a) Wähle im Anfangsschritt das volle Modell  $Y = \sum_{k=0}^K \beta_k x_k$
- b) In den weiteren Schritten wird jeweils eine Einflussgröße aus dem Modell genommen. Es wird jeweils die Einflussgröße, die zu dem höchsten  $R^2$  des resultierenden Modells führt, ausgeschlossen.
- c) **Stoppregel:** Nach bestimmtem Zielkriterium, z.B. AIC.

## 4 Schrittweise Selektion:

Kombination aus Vorwärts- und Rückwärtsselektion. Er wird eine Vorwärtsselektion und nach jedem Schritt eine Rückwärtsselektion mit geeignetem Stoppkriterium durchgeführt.

# Abschließende Bemerkungen

---

- Möglichst inhaltlich sparsames (Maximal-)–Modell vorgeben
- Stat. Tests **nach** Variablenselektion nicht durchführbar
- Strategie bei zentraler Einflussgröße
  - Variablenselektion für „Confounder“-Modell ohne zentrale Einflussgröße
  - Test auf Effekt der zentralen Einflussgröße mit dem gewählten Confounder-Modell
- Prognoseintervalle etc. mit Kreuzvalidierung (evtl. unter Einbeziehung der Selektion)

- Es wurden in den letzten Jahren Methoden entwickelt, die Modellschätzung und Variablenselektion simultan erreichen: LASSO und Boosting-Verfahren, siehe Vorlesung Generalisierte Regression bzw. Fahrmeier et al (2013).
- Es gibt in der Regel nicht einfach ein bestes Modell, sondern mehrere passende Modelle. Bayes- Verfahren ermöglichen Inferenz dazu durch Berechnung von Wahrscheinlichkeiten für bestimmte Modelle.
- Eine Strategie in der Inferenz und Prognose besteht in dem sog. Model- Averaging: Man bestimmt Prognosen aus mehreren Modellen als gewichtetes Mittel der Einzel-Prognosen.