Angewandte Statistik

Julian Hinz — Universität Bielefeld

Session 6

Modellwahl

Lernziel

Bisher

- *Parametrische* und *nichtparametrische* Regression

Heute

- Welches Modell ist das “richtige” Modell?

Modellwahl

Entscheidungen bei Regressionsanalyse:

- welche Variablen gehören ins Modell?
- brauche ich quadratische Terme?
- brauche ich Interaktionsterme?
- brauche ich ein nichtparametrisches Modell?

Wie kann man bei einer Vielzahl plausibler Modelle eine geeignete Wahl treffen?

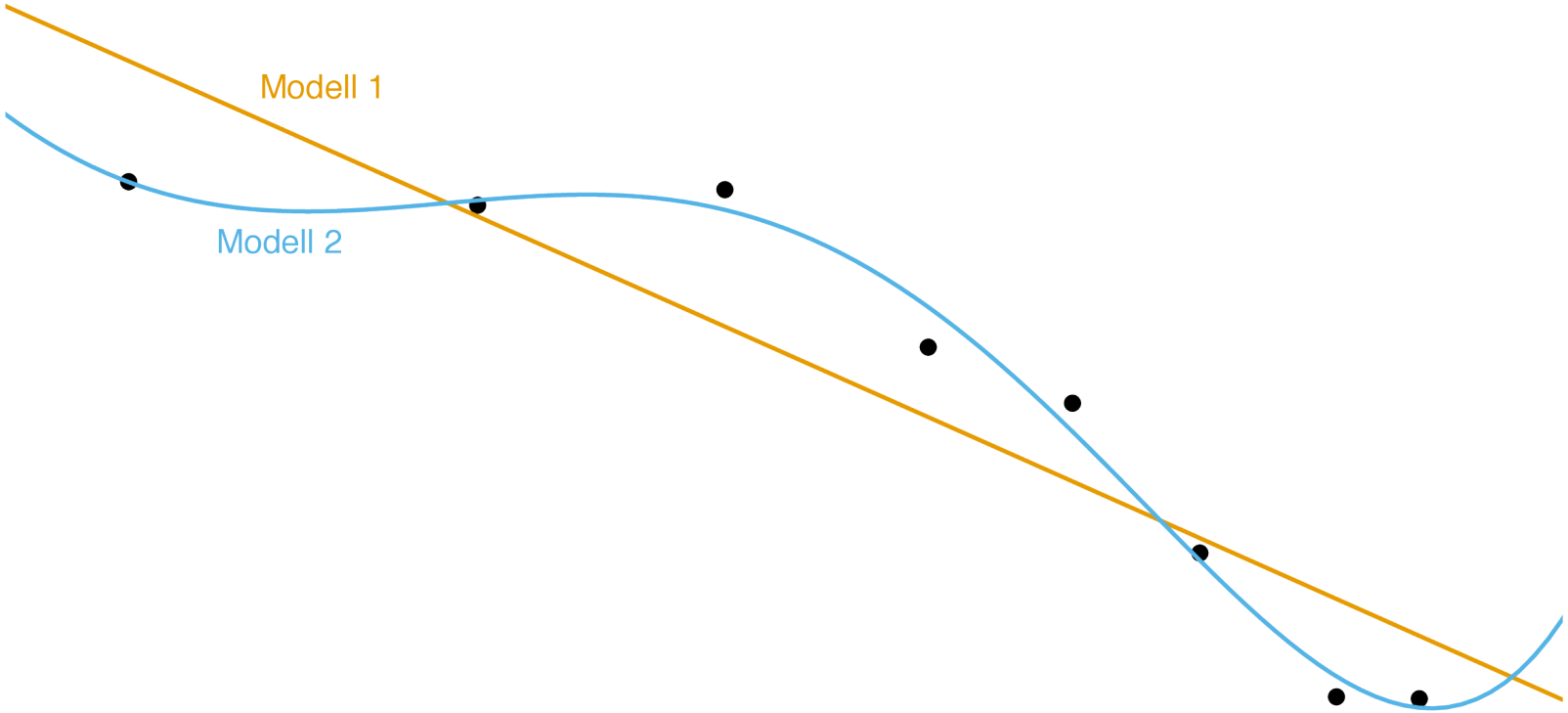
→ genau darum geht es bei der **Modellwahl** (oder auch: **Modellselektion**)

Ziele bei Modellwahl

Verschiedene Anwendungen und verschiedene Ziele:

- Vorhersage von Werten: Beispiel Fußball
- Zusammenhänge aufzeigen: Beispiel der Donner Party

Bias-Varianz Trade-off bei der Modellwahl



Bias-Varianz Trade-off bei der Modellwahl

Beispiel Donner Party Modell:

$$M_1 : \ \pi_i = h(\beta_0 + \beta_1 \cdot \text{alter}_i)$$

weniger Potenzial als das Modell

$$M_2 : \ \pi_i = h\Big(\beta_0 + \beta_1 \cdot \text{alter}_i + \beta_2 \cdot \text{alter}_i^2 + \beta_3 \cdot \text{geschlecht}_i + \beta_4 \cdot \text{fam.groesse}_i + \beta_5 \cdot \text{fam.groesse}_i^2 + \beta_6 \cdot \text{fam.groesse}_i \cdot \text{alter}_i\Big)$$

- Mit M_2 steigt aber das Risiko, zufälliges Rauschen zu modellieren: “Overfitting”

Simulationsexperiment Bias-Varianz Trade-offs

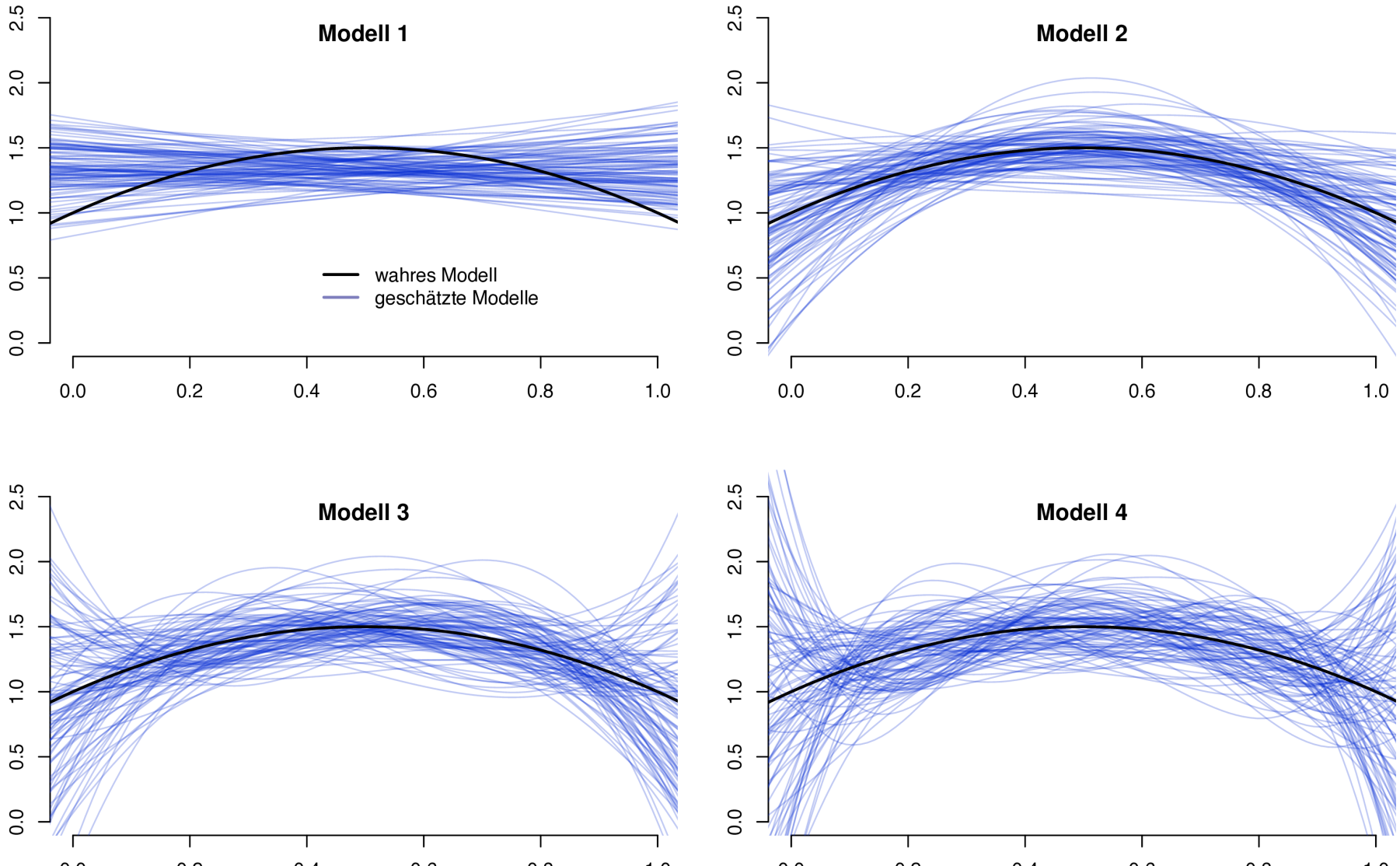
1. Simuliere 100 Datenpunkte aus $Y_i = 1 + 2x_i - 2x_i^2 + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$
2. Passe die folgenden linearen Modelle an:

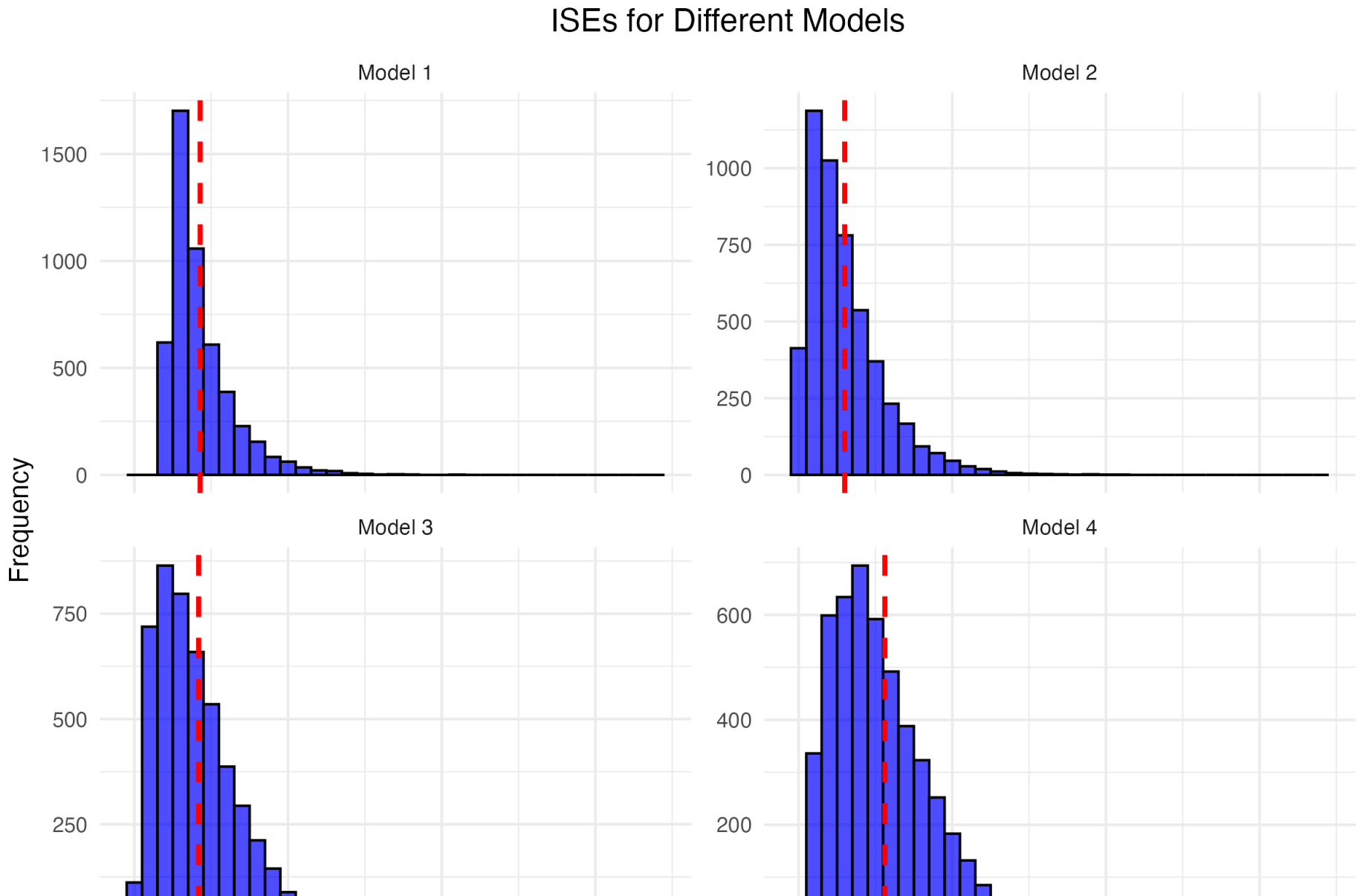
$$
\begin{aligned}
&\text{Modell 1:} && Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\
&\text{Modell 2:} && Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \\
&\text{Modell 3:} && Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i \\
&\text{Modell 4:} && Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_i
\end{aligned}
$$

3. Berechne für jedes dieser Modelle den integrierten quadrierten Fehler

$$\text{ISE} = \int \left(f(x) - \widehat{f}(x)\right)^2 dx,$$

wobei $f(x)$ die wahre und $\widehat{f}(x)$ die geschätzte Regressionsfunktion ist


Modell 1
Modell 2
Modell 3
Modell 4
wahres Modell
geschätzte Modelle
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0


ISEs for Different Models
Model 1
Model 2
Model 3
Model 4
Frequency
1500
1000
500
0
1000
750
500
250
0
750
500
250
600
400
200

Bedeutung des Bias-Varianz Trade-offs

- Modellwahl: richtige Balance zwischen Under- und Overfitting

Ein gutes Modell sollte:

- flexibel genug sein, um alle **systematischen Effekte** abzubilden …
  - sonst würde man systematisch daneben schätzen ⇝ hoher Bias!
- aber nicht zu flexibel sein, da sonst **zufällige Fehler** als systematische Effekte modelliert werden könnten
  - dann würde man zufällige Abweichungen/Rauschen ⇝ hohe Varianz!

also: Komplexität von Modell nur erhöhen, wenn es Sinn macht!

Modellwahl bei Regression

Problemstellungen bei gegebener Modellformulierung (z.B. Poissonregression):

1. geeignete Modellformulierung auswählen
   → z.B. lineares Modell vs. Variablentransformation vs. Splines …
2. wähle erklärende Variablen aus, die ins Modell aufgenommen werden sollen
3. entscheide ob polynomiale Terme und/oder Interaktionsterme aufgenommen werden sollen

Modellwahl im Falle der Regression

- für 2. und 3. können im Prinzip Hypothesentests genutzt werden
- z.B. bei linearer Regression kann man $H_0 : \beta_j = 0$ gegen $H_1 : \beta_j \neq 0$ testen
  - Ablehnung von $H_0 \rightsquigarrow$ Indikation, dass Einfluss vorhanden, daher drin lassen
  - Beibehaltung von $H_0 \rightsquigarrow$ Einflussgröße raus wg. Gefahr des Overfitting

Modellwahl bei mehreren möglichen Einflussgrößen

Iiteratives Verfahren, um Variablen auszuwählen:

- **Vorwärtsselektion**:
  - Beginne mit dem einfachsten Modell, $E(Y_i) = \beta_0$
  - nach und nach immer die “relevanteste” Variable aufnehmen…
  - …bis alle übriggebliebenen Variablen insignifikant sind
- **Rückwärtsselektion**:
  - Beginne mit dem Modell welches alle Variablen beinhaltet
  - nach und nach immer die am wenigsten relevante Variable entfernen…
  - …bis alle übriggebliebenen Variablen signifikant sind

Relevanz messen wir i.d.R. mit p-Werten (niedriger p-Wert = hohe Relevanz).

Wahl eines Modells im Beispiel der Donner Party

Rückwärtsselektion am Beispiel der Donner Party. Das komplexeste logistische Regressionsmodell, das wir betrachten, ist:

$$\begin{aligned}\pi_i = h\big(&\beta_0 + \beta_1 \cdot \text{alter}_i + \beta_2 \cdot \text{geschlecht}_i + \beta_3 \cdot \text{fam.groesse}_i \\ &+\beta_4 \cdot \text{alter}_i^2 + \beta_5 \cdot \text{fam.groesse}_i^2 + \beta_6 \cdot \text{alter}_i \cdot \text{fam.groesse}_i \\ &+\beta_7 \cdot \text{alter}_i \cdot \text{geschlecht}_i + \beta_8 \cdot \text{fam.groesse}_i \cdot \text{geschlecht}_i\big)\end{aligned}$$

(maximal quadratische Effekte & alle Interaktionen)

Modellwahl im Falle der linearen Regression - Reichen Hypothesentests?

- Durch Multikollinearität kann relevante Variable ausgeschlossen werden
- Kann nicht für alle Vergleiche herangezogen werden, z.B. nicht zum Vergleich der Modelle $Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$ und $Y_i = \beta_0 + \beta_1 \cdot \sqrt{x_i} + \epsilon_i$
- Nicht praktikabel, wenn eine große Zahl möglicher Modelle in Frage kommt
- Kann nicht auf nichtparametrische Modelle angewendet werden
- …

Likelihoodberechnung für Regressionsmodelle

Zur Erinnerung: Die **Likelihood**,

$$\mathcal{L}(\boldsymbol{\theta}) = f(y_1, \ldots, y_n | \boldsymbol{\theta}),$$

ist die gemeinsame Dichte der Beobachtungen $y_1, \ldots, y_n$ aufgefasst als Funktion des Parametervektors $\boldsymbol{\theta}$.

Im linearen Regressionsmodell,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

ist der Parametervektor $\boldsymbol{\theta} = \left(\beta_0, \beta_1, \sigma^2\right)$, die Beobachtungen $Y_1, \ldots, Y_n$ sind unabhängig voneinander und die Likelihood somit

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}\left(\beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} f_{N(\beta_0+\beta_1 x_i, \sigma^2)}(y_i).$$

Log-Likelihood im LSD-Beispiel

Log-Likelihood im LSD-Beispiel:

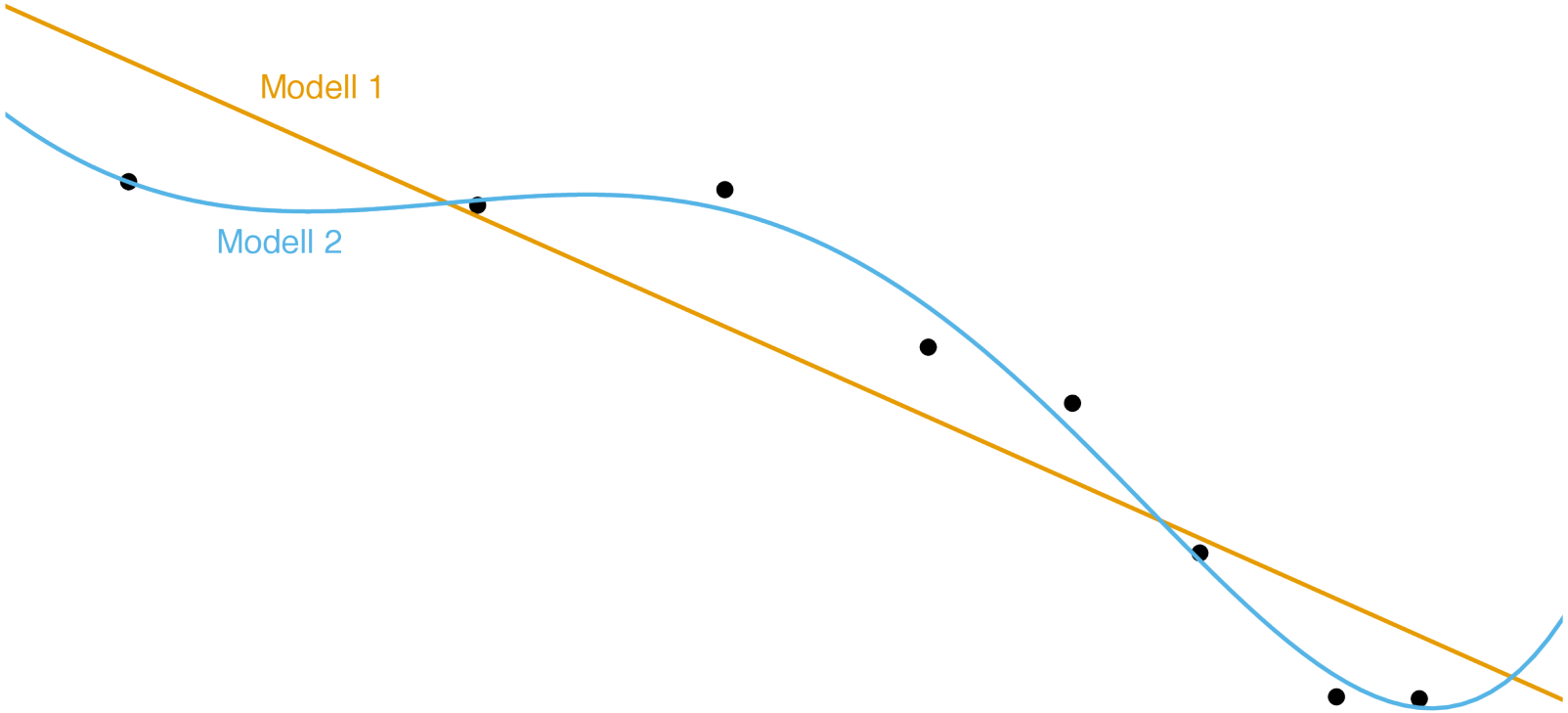
```r
LSD <- c(1.17, 2.97, 3.26, 4.69, 5.83, 6.00, 6.41)
Punkte <- c(78.93, 58.20, 67.47, 37.47, 45.65, 32.92, 29.97)
mod <- lm(Punkte ~ LSD)
logLik(mod)
# 'log Lik.' -22.50092
```

Likelihood als Kriterium zur Modellwahl?

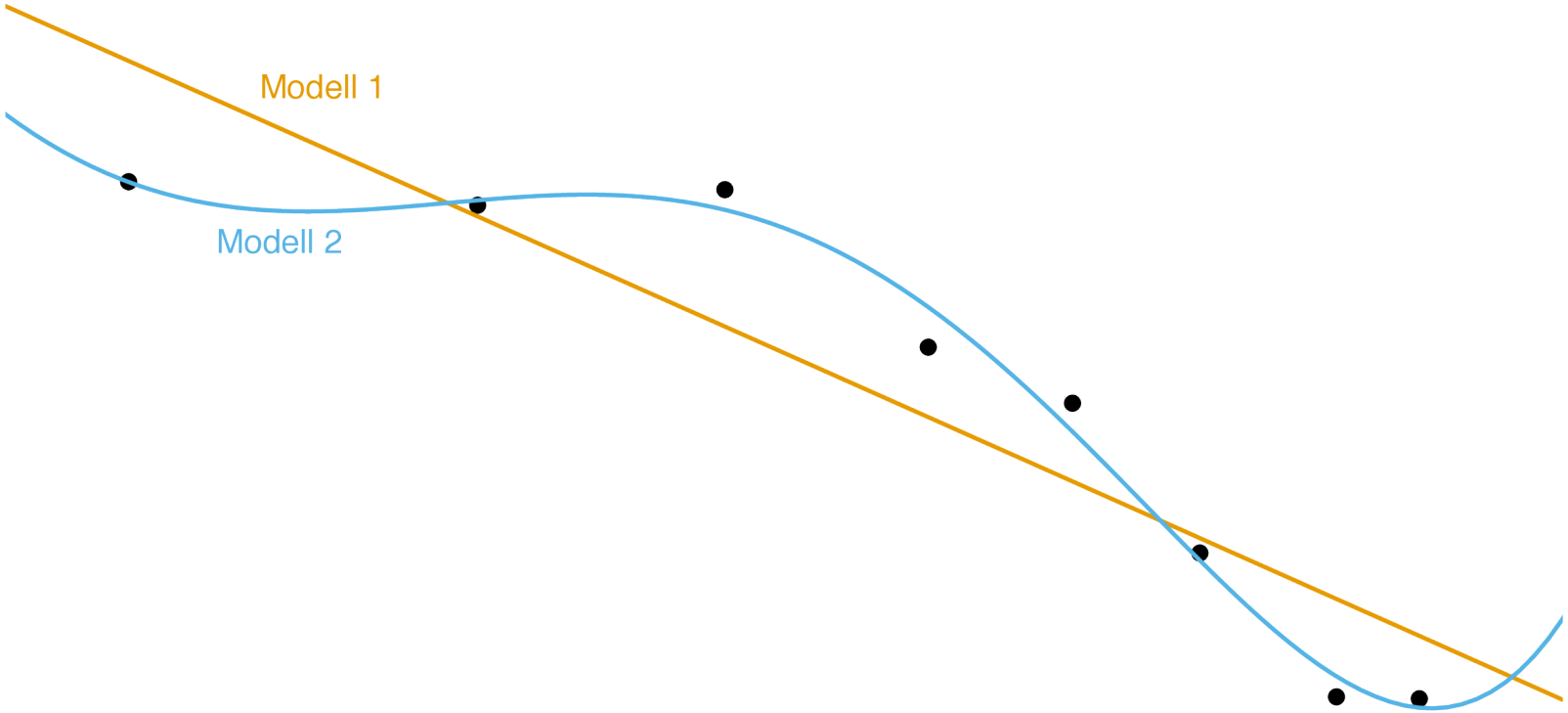
- Modell wählen das höchste Likelihood hat?
- Problem: mit höherer Komplexität wird die max. Likelihood niemals sinken
- Beispiel logistische Regression für Überlebenswahrscheinlichkeit bei Donner Party:

Modell	linearer Prädiktor	max. Log-Lik.
1	β_0	-58.62
2	$\beta_0 + \beta_1 \cdot \text{Alter}_i$	-55.12
3	$\beta_0 + \beta_1 \cdot \text{Alter}_i + \beta_2 \cdot \text{Alter}_i^2$	-52.53
4	$\beta_0 + \beta_1 \cdot \text{Alter}_i + \beta_2 \cdot \text{Alter}_i^2 + \beta_3 \cdot \text{Alter}_i^3$	-51.48

Likelihood als Kriterium zur Modellwahl?



Likelihood als Kriterium zur Modellwahl?



Kurze Bestandsaufnahme

Bei der Modellwahl geht es immer um den Bias-Varianz-Trade-off:

- brauchen hinreichend viel Flexibilität, um Struktur zu erfassen…
- …aber auch nicht mehr: Overfitting
- Likelihood misst Anpassungsgüte des Modells an die Daten
- Aber: Likelihood *allein* ist kein sinnvolles Kriterium, Komplexität nicht berücksichtigt

Akaike Information Criterion (AIC)

Akaike konnte 1974 zeigen, dass — unter einigen Annahmen — der Schätzfehler approximativ der Anzahl der Modellparameter entspricht.

- Wenn man den naiven Schätzer entsprechend korrigiert, erhält man

$$\log \mathcal{L}_{\hat{M}} - \text{Anz. Parameter}$$

  als zu maximierendes Kriterium.
- Kriterium wird meist mit -2 multipliziert:

$$-2 \log \mathcal{L}_{\hat{M}} + 2 \cdot \text{Anz. Parameter},$$

  eine Größe welche wir dann **minimieren** wollen.

Akaike Information Criterion (AIC) Formula

- **Akaike Information Criterion** (AIC) für gegebenes Modell ist

$$\mathrm{AIC} = -2 \log \mathcal{L}_{\widehat{M}} + 2 \cdot \text{Anz. Parameter},$$

wobei $\mathcal{L}_{\widehat{M}}$ der maximale Wert der Log-Likelihood ist - Modellen wählen wir jenes, welches das *niedrigste* AIC aufweist

- **belohnt Anpassungsgüte**: je näher Modell an Daten, desto niedriger ist (I)
- **bestraft Komplexität**: je komplexer das Modell, desto höher ist (II)

→ “bestes” Modell liegt irgendwo in der Mitte

Wahl eines Modells im Donner Party Beispiel

Das komplexeste logistische Regressionsmodell, das wir betrachten, ist:

$$\begin{aligned}\pi_i = h\big(&\beta_0 + \beta_1 \cdot \text{alter}_i + \beta_2 \cdot \text{geschlecht}_i + \beta_3 \cdot \text{fam.groesse}_i \\ &+\beta_4 \cdot \text{alter}_i^2 + \beta_5 \cdot \text{fam.groesse}_i^2 + \beta_6 \cdot \text{alter}_i \cdot \text{fam.groesse}_i \\ &+\beta_7 \cdot \text{alter}_i \cdot \text{geschlecht}_i + \beta_8 \cdot \text{fam.groesse}_i \cdot \text{geschlecht}_i\big)\end{aligned}$$

Wahl eines Modells im Donner Party Beispiel

- insgesamt 8 mögliche Variablen, also 2^8 = 256 Modellkandidaten

alt	ges	fam.g	alter²	fam.g²	alt:ges	alt:fam.g	ges:fam.g	AIC
✓	✓	✓	✓	✓		✓	✓	90.921
✓	✓	✓	✓	✓			✓	90.964
✓	✓	✓	✓	✓		✓		92.226
✓	✓	✓	✓	✓	✓		✓	92.821
✓	✓	✓	✓	✓	✓	✓	✓	92.847
✓	✓	✓	✓	✓				93.108

Bemerkungen zum AIC

- AIC wählt ein komplexeres Modell aus als Hypothesentest-basierte Rückwärtsselektion
- AIC neigt allgemein dazu, tendenziell eher zu komplexe Modelle zu wählen: Strafterm für Komplexität recht niedrig
- Alternative bietet **Bayesian Information Criterion** (BIC)

$$\text{BIC} = -2 \log \mathcal{L}_{\hat{M}} + \log(n) \cdot \text{Anz. Parameter},$$

  wobei n die Anzahl der Beobachtungen ist.
- Für $n \geq 8$ bestraft BIC Komplexität stärker als das AIC

BIC-basierte Modellwahl

Die besten Modelle gemäß BIC (ausgenommen jene, welche einen Interaktionsterm ohne dazugehörigen Haupteffekt beinhalten):

alt	ges	fam.g	alter²	fam.g²	alt:ges	alt:fam.g	ges:fam.g	BIC
✓		✓	✓	✓				107.627
✓	✓	✓	✓	✓				108.764
✓	✓	✓	✓	✓			✓	108.063
✓	✓	✓	✓	✓		✓		109.324
✓		✓	✓	✓		✓		110.198
✓	✓	✓	✓	✓		✓	✓	110.462

Bestes Modell gemäß BIC:

$$\pi_i = h\left(\beta_0 + \beta_1 \cdot \text{alter}_i + \beta_2 \cdot \text{fam.groesse}_i + \beta_3 \cdot \text{alter}_i^2 + \beta_4 \cdot \text{fam.groesse}_i^2\right)$$

Allgemeiner Einsatz von AIC & BIC

- bisher auf Regressionsmodelle konzentriert
- AIC und BIC aber anwendbar auf alle statistischen Modelle, solange

1. die Likelihood verfügbar ist und
2. die Modellparameter sinnvoll gezählt werden können.

→ für nichtparametrische Regressionsmodelle ist 2. nicht einfach

Modellwahl in der Praxis

- Alle Kriterien nur “approximativ optimal”
- keine universell anwendbare Strategie, die zuverlässig zu guten Ergebnissen führt: Wegweiser*

Modellwahl in der Praxis:

- Scharfes Nachdenken, welche Modelle konzeptuell sinnvoll sind
- Intuition und Pragmatismus
- Genaue Modellüberprüfung

Berühmte Zitate von George Box

> “Since all models are wrong the scientist cannot obtain a ‘correct’ one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.”

> “[…] there is no need to ask the question ‘Is the model true?’. If ‘truth’ is to be the ‘whole truth’ the answer must be ‘No’. The only question of interest is ’Is the model illuminating and useful?”

> “[…] Essentially, all models are wrong, but some are useful.”