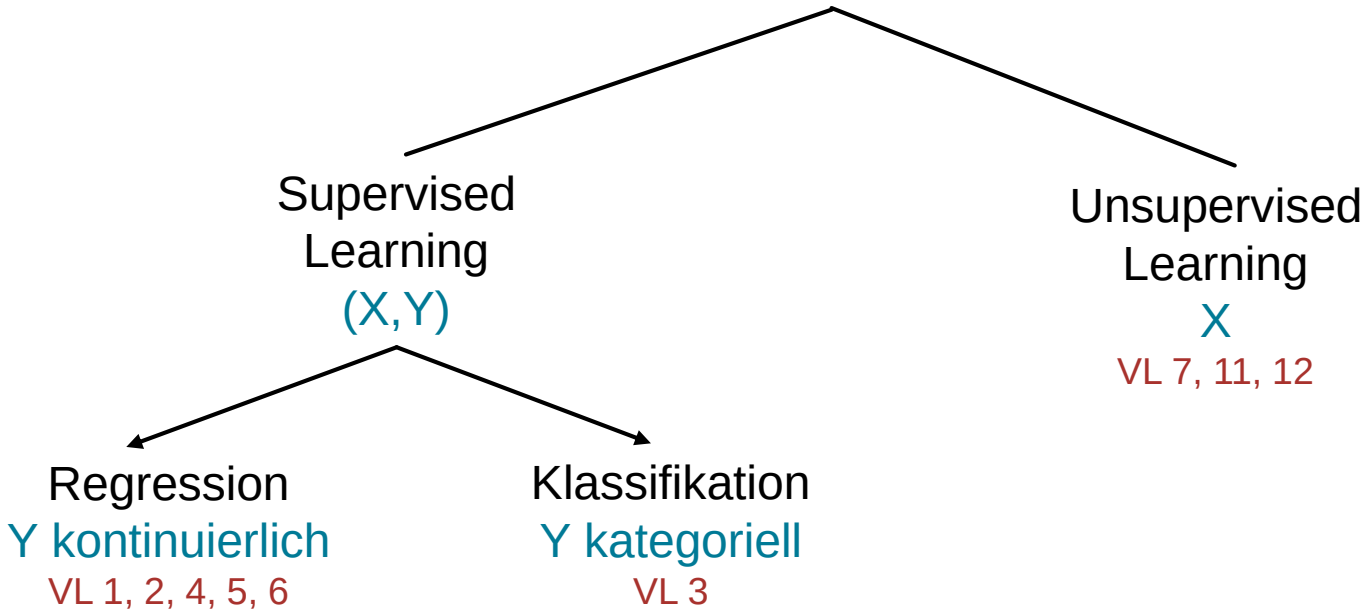
ETH zürich

Lineare Regression 2: Gute Vorhersagen

Big Picture: Statistisches Lernen



Exkurse: VL 8 (Poweranalyse), VL 9 & 10 (Experimentelles Design),
VL 13 (Fehlende Werte, Reproduzierbarkeit)

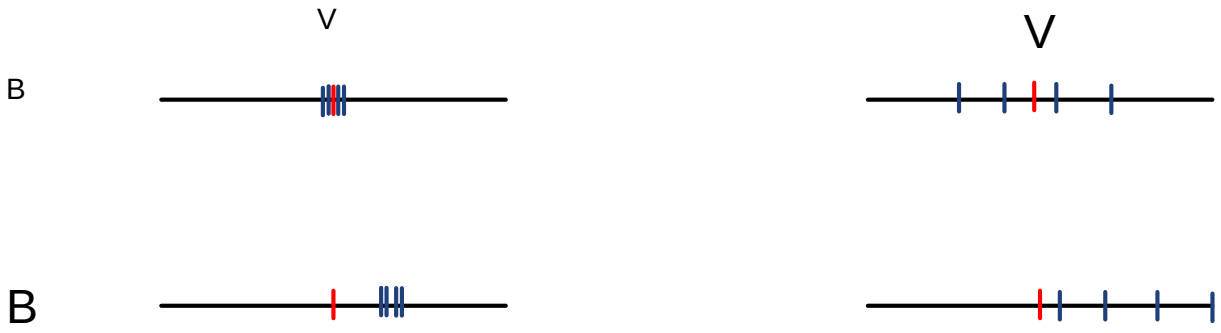
Ziele: Vorhersage vs. Inferenz

- Vorhersage: Hintergründe egal (Black box)
  - Gegeben die Blutwerte: Verträgt der Patient das Medikament ?
  - Wie viele Kunden wollen neues Produkt ?
  → Möglichst komplexes Modell, um alle Details zu erfassen
- Inferenz: Hintergründe verstehen
  - Durchschnittliche Kassierzeit pro Produkt ?
  - Gibt es einen Zshg zw Dosis und Heilungswa. ?
  → Möglichst einfaches Modell, um interpretieren zu können

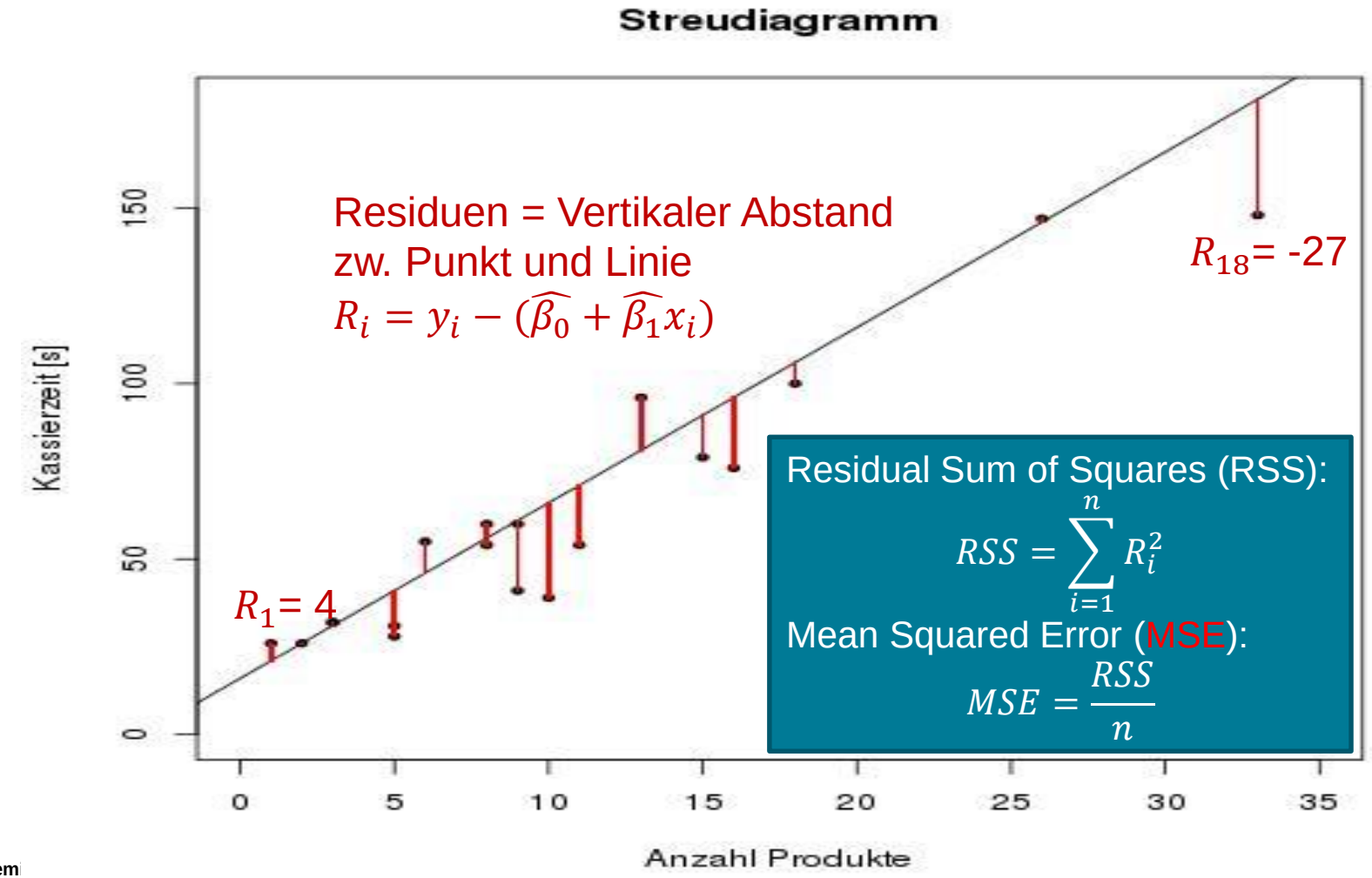
- Lineare Regression: Guter Kompromiss
  → werden uns weiter damit beschäftigen
- Fokus heute: Vorhersage

Wdh: Bias und Varianz eines Schätzers

- Schätzer $\widehat{\Theta}$, wahrer Parameter Θ
- Je nach beobachteten, zufälligen Daten: $\widehat{\Theta}$ variiert
- Bias (B) von $\widehat{\Theta}$: $E(\widehat{\Theta} - \Theta)$
- Varianz (V) von $\widehat{\Theta}$: $\mathrm{Var}(\widehat{\Theta}) = E\left((\widehat{\Theta} - \Theta)^2\right)$



Wdh: Güte vom Modell



Training MSE vs. Test MSE

- Training Daten: Bisher gesehene Daten
  Test Daten: Neue, zukünftige Daten
- Training MSE: Fehler auf bisher gesehenen Daten
  Test MSE: Fehler auf zukünftigen Daten
- Bisher: “Gesehenes gut erklären”
  - Modellklasse wählen (z.B. Geradengleichung)
  - Parameter finden, so dass Trainings MSE minimal
- Neues Ziel: “Zukünftige Daten gut erklären”
  Modell finden, so dass Test MSE minimal

Bias-Variance Trade-Off

- Training MSE: Kann beliebig klein gemacht werden, wenn wir nur genügend Parameter verwenden
- Test MSE: Selbst wenn wir $f(x)$ perfekt schätzen, stört der Fehlerterm ε unsere Vorhersage
- (Eq. 2.7): Erwartetes Residuenquadrat an Stelle x_0 im Testset:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\varepsilon)$$

  Test MSE: Mittelwert von $E\left(y_0 - \hat{f}(x_0)\right)^2$ über alle möglichen Werte von x_0 im Testset.
  → perfekter Fit von $f(x)$: $Test\ MSE = Var(\varepsilon) > 0$

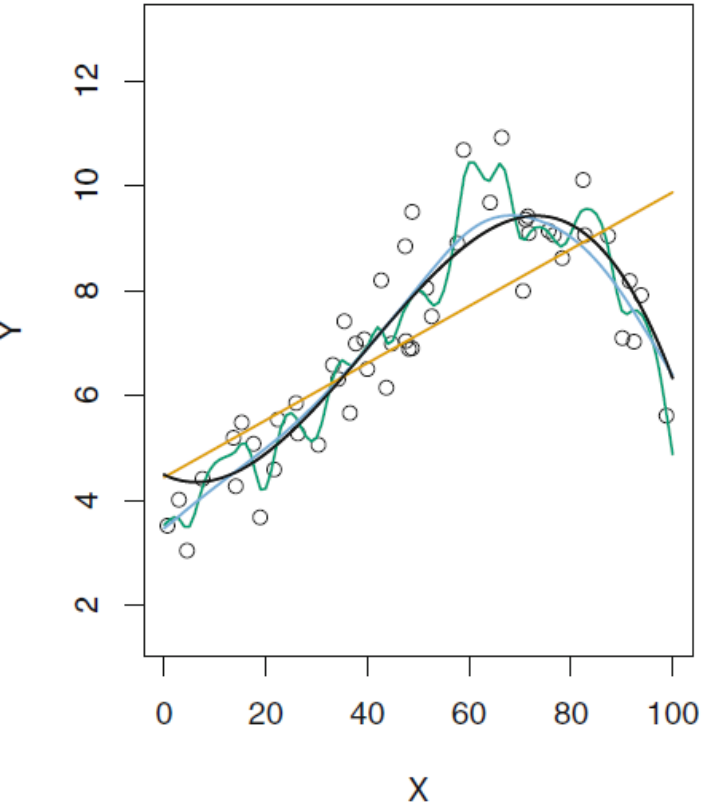
- Fazit: Training MSE ≠ Test MSE

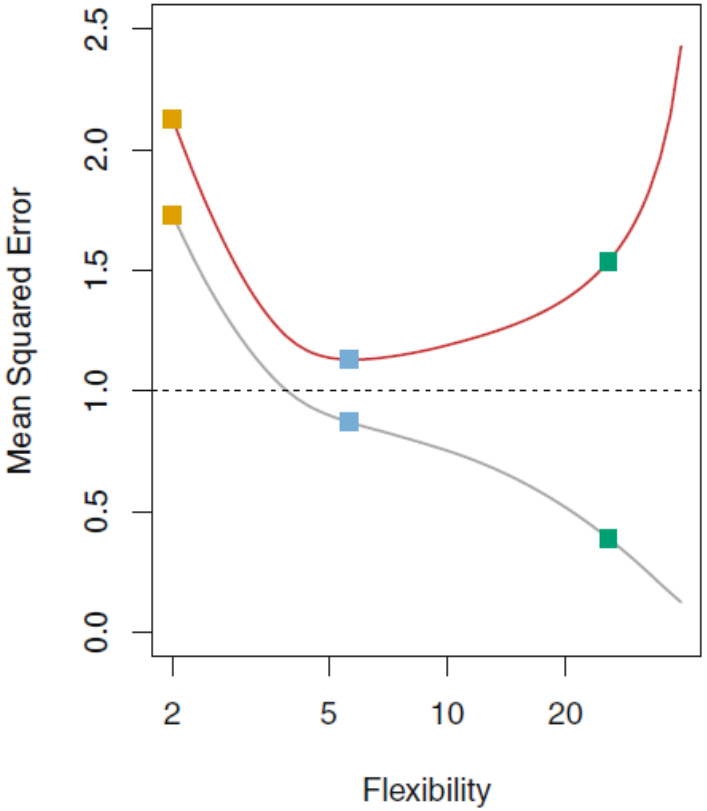
Training MSE ≠ Test MSE

$$Y = f(x) + \varepsilon$$

Schwarze Kurve: $f(x)$

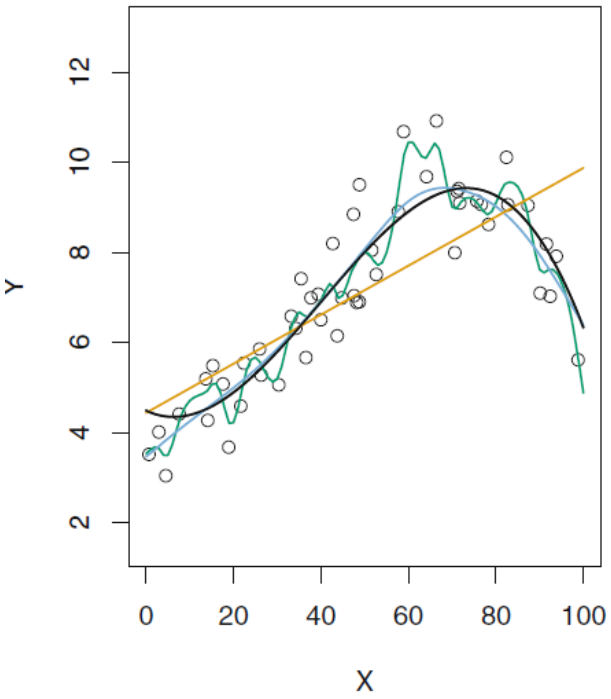
Welche Kurve beschreibt
- Training MSE
- Test MSE ?

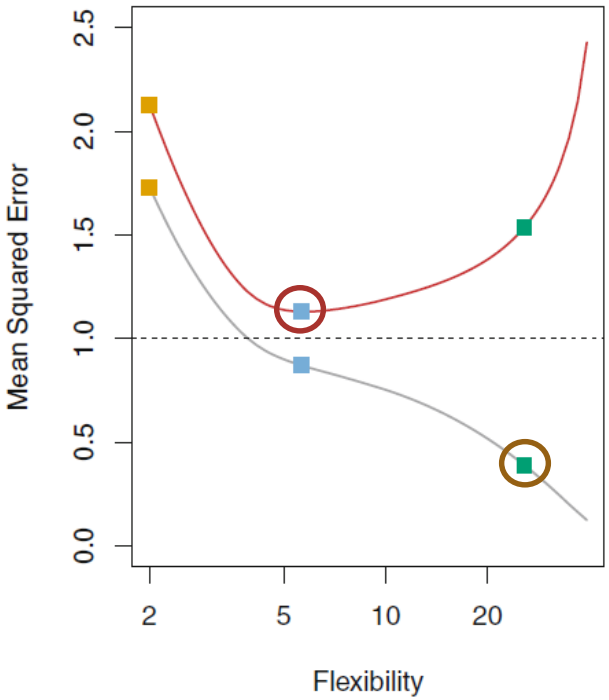




Paradox: “Overfitting”

- Paradox: Modell das die bisherigen Daten am besten beschreibt (minmaler Training MSE) ist nicht unbedingt das beste für zukünftige Daten (minimaler Test MSE)!





$Y = f(x) + \varepsilon$
$f(x)$ meist “glatt”

Perfekter Fit auf Trainings-Daten modelliert v.a. Fehlerterm, der bei zukünftigen Daten anders sein wird.

Fazit

- Um gute Vorhersagen zu machen, müssen wir den Test MSE minimieren

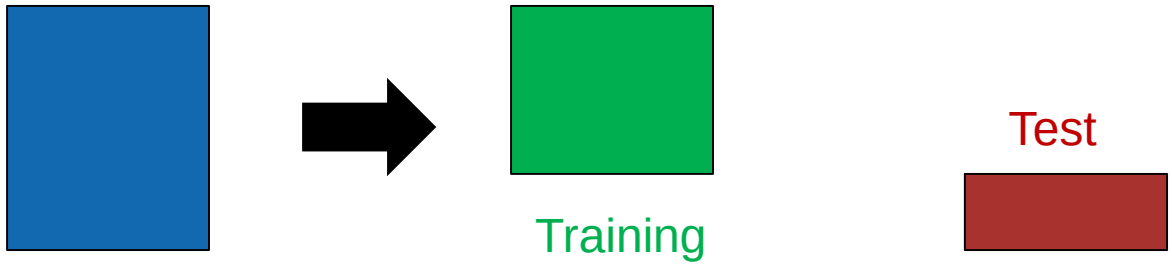
- Um den Test MSE zu minimieren reicht es NICHT den Training MSE zu minimieren

Wie schätzt man den Test MSE ?

- Direkte Methoden
  - Test Datensatz
  - Cross-validation (CV)
- Indirekte Methoden: $C_p, AIC, BIC, Adjusted\ R^2$
  - Korrektur vom Training MSE
  - Approximation von direkter Methode
  - Schnell: Gut, wenn viele oder aufwändige Modelle zu schätzen sind

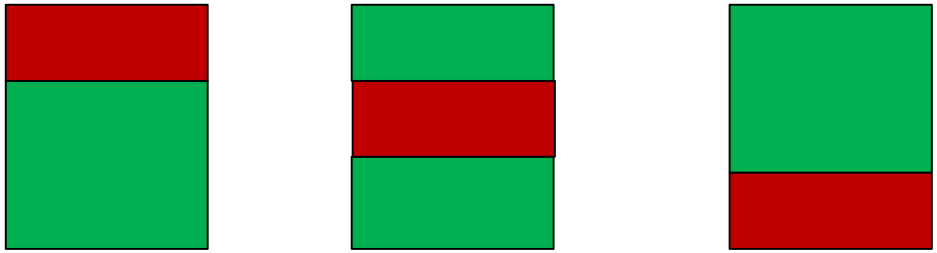
Direkte Methode 1: Expliziter Test Datensatz

- Teile Daten in Test- und Trainingsdatensatz



- Schätze Modell auf Trainingsdatensatz, evaluiere auf Testdatensatz
- Vorteil: Einfach, schnell
- Nachteil:
  - Je nach Wahl vom Testdatensatz: Unterschiedlicher MSE
  - Trainingsdatensatz kleiner als Originaldatensatz → Test MSE wird überschätzt

Direkte Methode 2: Cross-Validation (CV)



- Leave-one-out cross-validation (LOOCV):
  Jede Zeile ist einmal Testset; Rest ist Trainingsset
  Nachteil: Langsam, weil ein Fit pro Zeile
- K-fold cross-validation, z.B. 10-fold:
  Teile Daten in 10 Blöcke; jeder Block ist einmal Testset
  Nachteil: Je nach Unterteilung in K Blöcke unterschiedliches Test MSE

Cross-Validation in R

- **Grundsätzlich funktioniert CV für alle erdenklichen Vorhersagemethoden**
- Für Lineare Modelle (und sogar GLM's) gibt es eine besonders einfache Funktion: cv.glm() in package 'boot'
- Wdh: Lineare Modelle sind eine Unterklasse der Generalized Linear Models (GLMs)
- Damit kann man sowohl LOOCV als auch k-Fold CV durchführen

Indirekte Methoden: Hintergrund

- Kriterium K, das Güte vom Fit (RSS) und Anzahl Parameter (d) verbindet: $K = RSS + f(d)$
- Theorie: Unter gewissen Annahmen und bei sehr grossen Datensätzen (asymptotisch) gilt:
  Modell mit bestem K ist optimal für Vorhersage
- Vorteil: Schnell zu rechnen
- Nachteil: Approximativ; macht Annahmen; Test MSE nicht berechnet
- Praxis: Verwenden, falls viele oder komplizierte Modelle geschätzt werden müssen

Indirekte Methoden: Überblick

Je nach theo. Annahmen, leicht andere Form von K
(d: Anz. Parameter im Modell, n: Anz. Beobachtungen)

- $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$
- $AIC = \frac{1}{n\,\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$
- $BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$
- Adjusted $R^2 = 1 - \left(\frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}}\right)$ (TSS: Total sum of squares)

- Praxis: Willkürlich für eine Methode entscheiden
  (z.B. BIC)

Modellwahl

- Fokus 1: Finde Modell, das möglichst kleinen Vorhersagefehler hat
- Fokus 2: Finde Variablen, die für gute Vorhersage nötig sind

- Könnten CV verwenden (s. ISL 6.5.3); wir konzentrieren uns aber auf indirekte Methoden: Einfacher und schneller

- Faustregel für die Praxis:
  - Test MSE mit CV (direkte Methode)
  - Modellwahl mit BIC (indirekte Methode)

Bsp: Gehalt von Baseball Spielern

- Erkläre Gehalt (“Salary”) von Baseball Spielern mit erklärenden Variablen in der Saison 86/87
- Datensatz “Hitters” im package “ISLR”

```
> names(Hitters)
 [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"     "Years"     "CAtBat"
 [9] "CHits"     "CHmRun"    "CRuns"     "CRBI"      "CWalks"    "League"    "Division"  "PutOuts"
[17] "Assists"   "Errors"    "Salary"    "NewLeague"
```

Techniken zur Modellwahl 1: Exakt

- Berechne eine Lineare Regression für alle möglichen Kombinationen von erklärenden Variablen; speichere BIC
- Vorteil: Findet bestes Subset bzgl. BIC
- Nachteil: Rechenaufwand !
  p Variablen → 2^p Subsets

p	2^p
10	10^3
20	10^6
30	10^9
40	10^{12}
…	…

In der Praxis kaum mehr zu berechnen

Techniken zur Modellwahl 2: Heuristiken

- Wie Bergwanderung im Nebel: Gehe immer nach oben → man landet evtl. auf Zwischengipfel
- Verfehlen evtl. globales Optimum

- “Stepwise forward”: Starte mit dem leeren Modell; füge immer eine Variable hinzu
- “Stepwise backward”: Start mit dem vollen Modell; lasse immer eine Variable weg

RSS oder BIC ?

Angenommen, wir haben 20 Variablen zur Auswahl. Wir wollen das beste (bzgl. Vorhersage) Modell mit genau 5 Variablen finden. Sollten wir als Gütekriterium RSS oder BIC verwenden?

- RSS produziert das bessere Modell.
- BIC produziert das bessere Modell.
- RSS und BIC produzieren das gleiche Modell.

Bsp: Stepwise forward selection

- Variablen: Y, X1, X2, X3
  M1: $Y \sim 1 \rightarrow BIC = 20$
- Bestes Modell mit einer Variable:
  M2: $Y \sim X2 \rightarrow BIC = 17$, also besser
- Bestes Modell mit $X2$ und noch einer Variable:
  M3: $Y \sim X2 + X1 \rightarrow BIC = 18$,
  also schlechter als $M2 \rightarrow$ Stop

- Ausgabe: Bestes Modell ist $Y \sim X2$.

Bsp: Stepwise backward selection

- Variablen: Y, X1, X2, X3

  M1: $Y \sim X1 + X2 + X3 \rightarrow BIC = 20$
- Bestes Modell mit einer Variable weniger:
  M2: $Y \sim X2 + X3 \rightarrow BIC = 17$, also besser als $M1$
- Bestes Modell mit einer Variable weniger als $X2, X3$:
  M3: $Y \sim X2 \rightarrow BIC = 18$, also schlechter als $M2 \rightarrow$ Stop

- Ausgabe: Bestes Modell ist $Y \sim X2$.

Modellwahl in R

- Funktion “regsubsets” in Paket “leaps”;
  berechnet sowohl exakt als auch mit Heuristiken
- Definition von BIC in “leaps”:

$$BIC = -\frac{1}{n}(RSS + \log(n)\, d\hat{\sigma}^2)$$

  → finde Modell mit minimalem BIC
- Vorgehen:
  1) Für jede Anzahl erklärende Variablen: Finde bestes Subset bzgl. RSS

```
m1 <- regsubsets(Salary ~ ., data = Hitters)
```

  2) Vergleiche Modelle mit unterschiedlichen Variablenzahlen mit BIC

```
which.min(m1s$bic)
```

- Bemerkung: Falls Anzahl Variablen fix ist, finden RSS und BIC das gleiche optimale Modell