Bootstrap Model Averaging zur Berücksichtigung von Modellselektionsunsicherheit im linearen Regressionsmodell

Norbert Holländer[1], Nicole Augustin[2] & Willi Sauerbrei[1]

[1] Institut für Medizinische Biometrie und Medizinische Informatik, Universitätsklinikum Freiburg

[2] Department of Statistics, University of Glasgow

Erlangen, 23. Januar 2004

Überblick

- Modell und Variablenselektion
- Model Averaging: Ein Bootstrap-Ansatz
- Simulationsstudie
  - ⋆ Design
  - ⋆ Bewertungskriterien
  - ⋆ Resultate
- Diskussion

Lineares Regressionsmodell

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_p X_p + \epsilon = \boldsymbol{X\beta} + \epsilon$$

- $Y = (Y_1, ..., Y_n)'$ Responsevektor
- $X_1, ..., X_p$ Vektoren der (potentiellen) Einflussgrößen
- $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ mit $\epsilon \sim N(0, \sigma^2 I)$

Modellselektion

- Kategorisierung / Bestimmung der funktionalen Form einzelner X_i
- Variablenselektion (z.B. backward elimination BE)

Probleme

- Selektionsprozess bei der Parameterschätzung nicht berücksichtigt
- viele Faktoren → viele potentielle Modelle
  einfacher Fall: Kovariable in/out
  $p = 10 \rightarrow 2^{10} = 1024$ Modelle
  $p = 25 \rightarrow 2^{25} = > 33$ Mio Modelle
- ein selektiertes Modell, aber viele Modelle u.U. gleich gut
  → Modellselektionsunsicherheit wird ignoriert

Ziel von Model Averaging Methoden

Berücksichtigung der Modellselektionsunsicherheit

Methoden zur Modellselektion

- Volles Modell (alle Kovariablen)

- Backward Elimination mit AIC- oder BIC-Kriterium (BE_{AIC}, BE_{BIC})
  $\rightarrow$ möglichst guter Fit, aber Bestrafungsterm für Anzahl der Kovariablen ($AIC = 2\,\tilde{p}$, $BIC = \log(n)\,\tilde{p}$)

Model averaging: Ein Bootstrap-Ansatz (Buckland et al, 1997; Augustin et al, 2002)

Generiere B Bootstrap Stichproben

Modellselektion: Modelle M_j, j=1,..,B

interessierende Zielgröße $\theta = E(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$

Schätzung mit Model Averaging:

$$\hat{\theta} = \sum_j \hat{\theta}_j \hat{w}_j$$

hier: $\hat{w}_j = h(M_j)$ - Anteil der Bootstrap-Stichproben mit selektiertem Modell M_j.

Bootstrap Model averaging (bootstrapMA) - Modellselektion in 2 Schritten

1. Schritt (Screening)

- Backward elimination (BE_{AIC}) in allen Bootstrap-Stichproben
- ermittle die Aufnahmehäufigkeiten $h(x_i)$
- verwende nur Einflussfaktoren X_i mit $h(x_i) > C$ (z.B. C=30%)

2. Schritt

- Backward elimination (z.B. BE_{AIC}) in neuen Bootstrap-Stichproben
- ermittle die Aufnahmehäufigkeiten $h(M_j)$
- berechne

$$\hat{\theta} = \sum_j \hat{\theta}_j \; h(M_j)$$

wobei $\hat{\theta}_j$ in den Originaldaten geschätzt wird

Warum Screening im 1. Schritt?

- **wichtig, wenn die Anzahl potentieller Kovariablen groß ist**

$\rightarrow$ 'unwichtige' Kovariablen sollen ganz eliminiert werden

$\rightarrow$ Schätzer $\hat{w}_j = h(M_j)$ für relevante Modelle M_j 'sicherer'

Ein Beispiel

Kovariablen	1. Schritt $h(x_i)$ %	2. Schritt				
X_1	20	-	-	-	-	-
X_2	27	-	-	-	-	-
X_3	18	-	-	-	-	-
X_4	31	-	x	-	x	x
X_5	99	x	x	x	x	x
X_6	100	x	x	x	x	-
X_7	99	x	x	-	-	x
X_8	100	x	x	x	x	x
X_9	100	x	x	x	x	x
X_{10}	100	x	x	x	x	x
$h(M_j)$ %	-	70.2	27.0	1.6	0.8	0.4

Ähnlich: bootstrap averaging (bagging) (Breiman, 1996)

$$\hat{\theta} = \frac{1}{B}\sum_{j}^{B} \hat{\theta}_j^b$$

hier: $\hat{\theta}_j^b$ geschätzte Zielgröße in der j-ten Bootstrap-Stichprobe basierend auf Modell M_j

vorher: $\hat{\theta}_j$ in den Originaldaten geschätzt,
d.h. für $M_j = M_k$ ist $\hat{\theta}_j = \hat{\theta}_k$

Simulation: Design (1)

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_p X_p + \epsilon = \boldsymbol{X\beta} + \epsilon$$

mit $(X_1, X_2, X_3, ..., X_p)$ multivariat normalverteilt, $\epsilon \sim N(0, \sigma_e^2 I)$
($\sigma_e^2 = 1.00 \longrightarrow$ adj. $R^2 \approx 0.85$ $\quad \sigma_e^2 = 6.25 \longrightarrow$ adj. $R^2 \approx 0.50$)

n = 100, B = 500 Bootstrap-Stichproben, R=100 Replikationen

Modell 1 (10 Kovariablen)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$\boldsymbol{\beta}^{true}$	0	0	0	-0.5	0.5	0.5	0.5	1	1	1.5

Korrelationen: $r_{X1,X5} = 0.7$ $\quad r_{X2,X6} = 0.5$ $\quad r_{X7,X8} = 0.3$
$r_{X1,X10} = 0.5$ $\quad r_{X4,X8} = -0.7$

Simulation: Design (2)

Modell 2 (15 Kovariablen)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
$\boldsymbol{\beta}^{true}$	0	0	0	-0.5	0.5	0.5	0.5	1	1	1.5	0	0	0	0	0

Korrelationen: $r_{X1,X5} = 0.7$ $\quad r_{X2,X6} = 0.5$ $\quad r_{X7,X8} = 0.3$ $\quad r_{X9,X13} = 0.5$
$r_{X1,X10} = 0.5$ $\quad r_{X4,X8} = -0.7$ $\quad r_{X11,X12} = 0.7$ $\quad r_{X7,X14} = 0.5$

Modell 3 (25 Kovariablen)

Modell 2 + noise
(10 weitere X_i mit $\beta_i = 0$, keine Korrelation)

Simulation: Methoden zur Modellselektion

- Volles Modell (alle Kovariablen)
- Backward Elimination mit AIC- oder BIC-Kriterium (BE_{AIC}, BE_{BIC})
  → möglichst guter Fit, aber Bestrafungsterm für Anzahl der Kovariablen ($AIC = 2\,\tilde{p}$, $BIC = \log(n)\,\tilde{p}$)

Model Averaging Methoden

- bootstrapMA
  1. Schritt: Screening BE_{AIC} — C = 0, 20, 30 %
  2. Schritt: BE_{AIC}
  1. Schritt: C = 0, 20, 30 %
  2. Schritt: BE_{BIC}
- bagging
  BE_{AIC}: bagg.AIC
  BE_{BIC}: bagg.BIC

Simulation: Bewertung der Resultate

Daten zur Modellbildung: $D^{build} = (Y^{build}, \boldsymbol{X}^{build})$

Validierungsdaten: $D^{test} = (Y^{test}, \boldsymbol{X}^{test})$

Zielgröße: $\boldsymbol{\mu}^{test} = E(Y^{test} \mid \boldsymbol{X}^{test}) = \boldsymbol{X}^{test}\boldsymbol{\beta}^{true}$

$$\boldsymbol{\mu}^{test} = (\mu_1^{test}, ..., \mu_n^{test})'$$

Bewertungskriterien

- Mean squared error: $MSE = \frac{1}{n}\sum_{i=1}^{n}(\mu_i^{test} - \hat{\theta}_i)^2$
- Überdeckungswahrscheinlichkeit der Konfidenzintervalle (Coverage): Anteil der μ_i^{test} mit $\mu_i^{test} \in$ 95% Konfidenzintervall

Berechnung der Konfidenzintervalle für $\mu = \boldsymbol{X}^{test}\beta^{true}$

Einzelnes Modell (volles Modell, BE_{AIC}, BE_{BIC}):

Vorhersage: $\hat{\theta} = \boldsymbol{X}^{test}\hat{\boldsymbol{\beta}}^{build}$

Varianzschätzer: $\text{Var}(\hat{\theta}) = \sigma^2\,\boldsymbol{X}^{test}(\boldsymbol{X}'^{build}\boldsymbol{X}^{build})^{-1}\boldsymbol{X}'^{test}$

bzw. $\text{Var}(\hat{\theta}_i) = \sigma^2\boldsymbol{x}_i^{test}(\boldsymbol{X}'^{build}\boldsymbol{X}^{build})^{-1}\boldsymbol{x}_i'^{test}$

Varianzschätzer aus selektiertem Modell
→ Modellselektionsunsicherheit nicht berücksichtigt

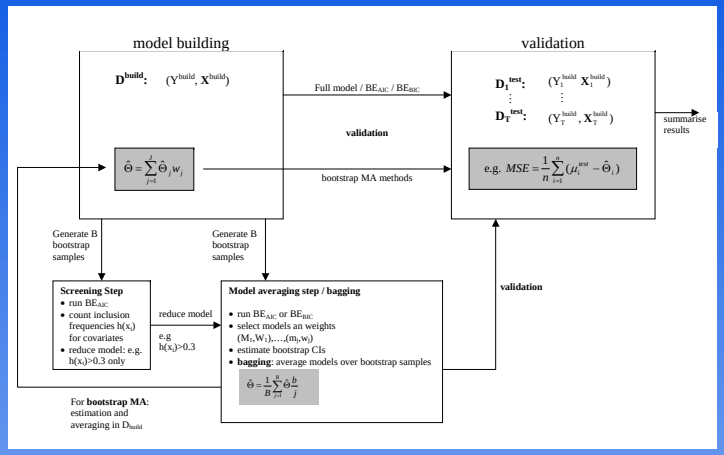
Berechnung der Konfidenzintervalle für $\mu = \boldsymbol{X}^{test}\beta^{true}$

Model averaging (boostrapMA bzw. bagging):

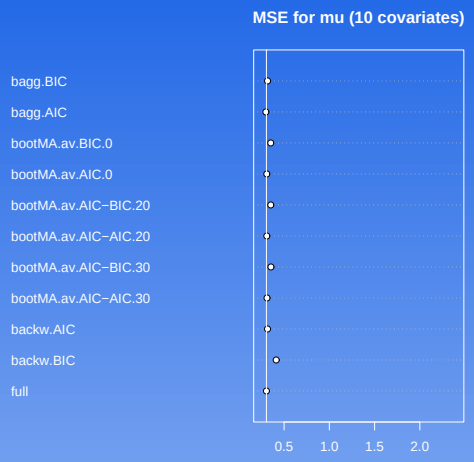
- Konfidenzintervall aus Bootstrap-Perzentilen
- Für boostrapMA: Varianzschätzer aus Buckland et al. (1997):

$$var(\hat{\theta}) = \left\{\sum_{j}^{J} w_j\sqrt{var(\hat{\theta}_j|M_j) + (\theta_j - \theta)^2}\right\}^2$$

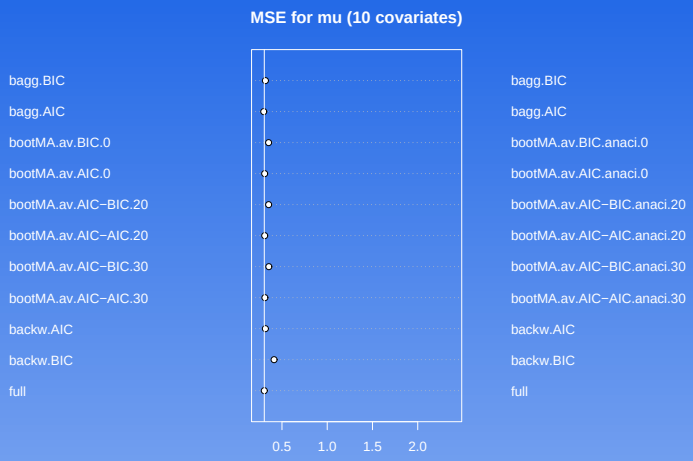
Unsicherheit der Modellselektion wird durch den *model misspecification bias* $(\theta_j - \theta)$ **berücksichtigt.**

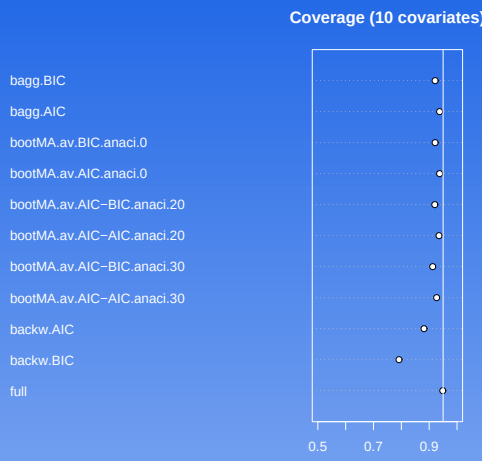


Modell 1 (10 Kovariablen)

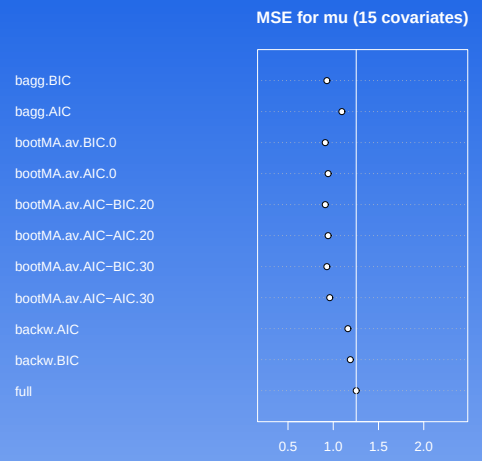


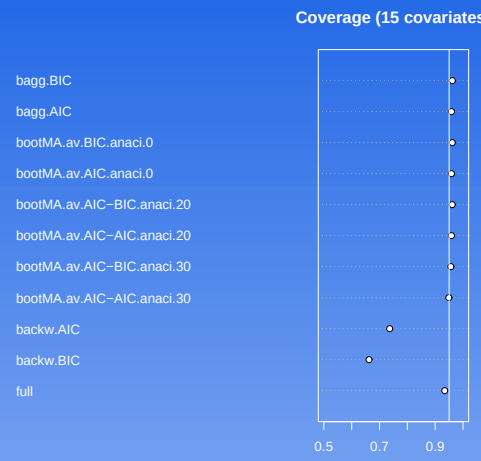
Modell 1 (10 Kovariablen)



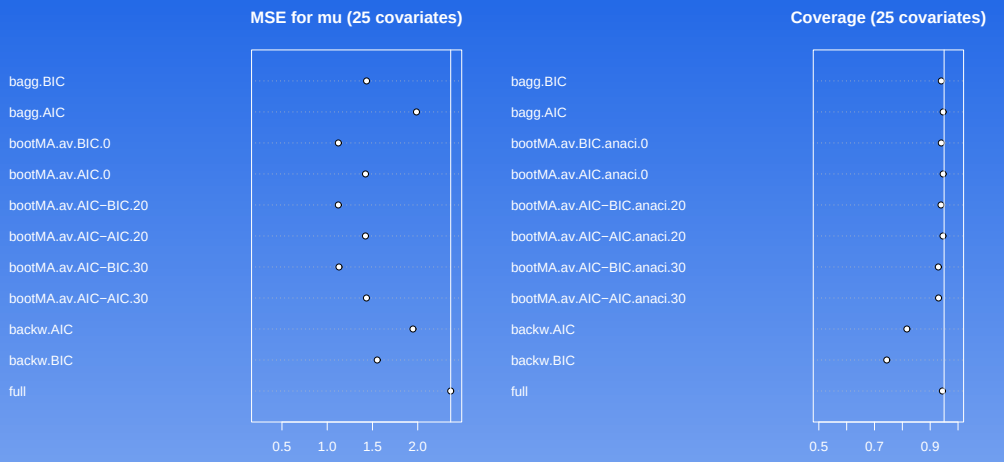


Modell 2 (15 Kovariablen)





Modell 3 (25 Kovariablen)



Zusammenfassung der Resultate

Modell	Anzahl Kovariablen		MSE	MSE im Vergleich zum vollen Modell BE / bootrapMA / bagging	Coverage der Konfidenzintervalle volles Modell BE	Coverage der Konfidenzintervalle volles Modell bootrapMA / bagging
1	10	1.7	klein	etwa gleich	unterschätzt	korrekt
2	15	2.6	mittel	etwas kleiner	unterschätzt	korrekt
3	25	4.3	groß	deutlich kleiner	unterschätzt	korrekt

mittlere Anzahl der im Screening-Schritt (C=30%) eliminierten Kovariablen

Diskussion (1)

- Modellselektionsunsicherheit besteht besonders dann, wenn viele Kovariablen keinen Effekt auf die Zielgröße haben
  $\longrightarrow$ bedingter Varianzschätzer aus einem selektierten Modell ignoriert Modellselektionsunsicherheit
- erste analytische Ergebnisse (Candolo et al. (2003), Hjort & Claeskens, 2003)
  $\longrightarrow$ Anwendbarkeit falls p groß ist ?
- Schwerpunkt der Forschung basiert auf Bayes-Methoden
  (Hoeting et al., 1999)

Diskussion (2)

- MSE nach Standard-Selektionsverfahren (z.B BE) häufig kleiner als im vollen Modell, aber Coverage von Konfidenzintervallen unterschätzt
- Bootstrap Model Averaging Methoden verkleinern MSE und liefern korrekte Coverage
- Screening-Schritt zur Elimination unwichtiger Variablen wichtig, um brauchbare Modelle zu erhalten
- erste Ergebnisse für das logistische Regressionsmodell deuten auf vergleichbare Resultate

Literatur

Augustin, N. H., Sauerbrei, W., & Schumacher, M. (2002):
Incorporating model selection uncertainty into prognostic factor model predictions.
FDM-Preprint Nr. 76, University of Freiburg. http://www.fdm.uni-freiburg.de/

Breiman, L. (1996): Bagging predictors. Machine Learning, 123–140.

Buckland, S.T., Burnham, K.P. & Augustin, N.H. (1997):
Model selection: an integral part of inference. Biometrics, 53, 603–618.

Candolo, C., Davison, A.C. & Demetrio C.G.B. (2003):
A note on model uncertainty in linear regression, The Statistician, 52, 165-177.

Hjort, N.L. & Claeskens, G. (2003): Frequentist model average estimators.
Journal of the American Statistical Association, 98, 879-899 (with discussion).

Hoeting , J.A., Madigan, D., Rafferty , A.E. & Volinsky, C.T. (1999):
Bayesian model averaging: A tutorial. Statistical Science, 14, 382-417.