Vergleich von Kriterien und Verfahren zur Modellwahl bei der multiplen linearen Regression

Joachim Spilke
Universität Halle-Wittenberg
Landwirtschaftliche Fakultät
06099 Halle
joachim.spilke@landw.uni-halle.de

Norbert Mielenz
Universität Halle-Wittenberg
Landwirtschaftliche Fakultät
06099 Halle
norbert.mielenz@landw.uni-halle.de

Zusammenfassung

Mit dem Ziel einer Bewertung der in der SAS-Prozedur Reg verfügbaren Kriterien zur Modellbildung erfolgt bei Nutzung der Monte-Carlo Simulation ein Vergleich der Vorhersagefehler. Die Ergebnisse zeigen eine deutliche Abhängigkeit der Wirksamkeit der Kriterien von Datenumfang und Modellkomplexität. Das korrigierte Akaike-Kriterium erweist sich als ein durchgehend nutzbares Kriterium. Bei einem geringem Datenumfang liefert jedoch die backward selection und stepwise selection einen geringeren Vorhersagefehler.

Schlüsselworte: multiple lineare Regression, Modellselektion, Vorhersagefehler.

1 Einleitung und Zielstellung

Die Regressionsanalyse stellt auch für biologische Anwendungen ein vielfach verwendetes statistisches Verfahren dar. Die breite Nutzung im biologischen Bereich erklärt sich aus den häufig auftretenden charakteristischen Anwendungsgebieten, die in der Ermittlung optimaler Werte von Einflussgrößen, der Korrektur von quantitativen Störgrößen oder der Merkmalsvorhersage bestehen (Draper and Smith, 1981 p. 5ff; Wetherill, 1985 p. 2ff; Rasch et al., 1996 p. 57ff).

Nachfolgend beschränken wir uns auf lineare Zusammenhänge. D.h., es wird als bekannt vorausgesetzt, dass die Zusammenhänge zwischen Einflussgrößen und der Zielgröße durch eine lineare Funktion adäquat abgebildet werden können. Bereits bei Beschränkung auf einen linearen Funktionstyp besteht eine wichtige Aufgabenstellung für den Anwender darin zu entscheiden, welche Merkmale als Einflussgrößen in das Modell aufzunehmen sind. Mit der damit verbundenen Modellwahl soll einerseits vermieden werden, Merkmale mit unbedeutendem Einfluss auf die Zielgröße einzubeziehen und den ggf. damit verbundenen Aufwand zur Merkmalserfassung unnötig zu betreiben. Andererseits soll damit aber auch die Gefahr gemildert werden, für die Beschreibung der Zielgröße bedeutsame Merkmale zu übersehen.

Aus der Literatur sind verschieden Kriterien zur Unterstützung der Modellwahl bekannt und auch in entsprechende Softwareprodukte umgesetzt, wie beispielsweise in SAS (Proc Reg). Es fehlt jedoch an Aussagen zur relativen Vorzüglichkeit der Kriterien und darauf aufbauenden begründeten Empfehlungen für den Nutzer, welches der angebotenen Kriterien verwendet werden soll. Nachfolgend sollen verschiedene

Kriterien zur Unterstützung der Modellwahl auf Basis der stochastischen Simulation vergleichend untersucht werden. Als Kriterium wird der mittlere Vorhersagefehler PMSE (Burnham and Anderson, 2004) verwendet, da damit der Zielstellung der Nutzung der Modellbildung und Schätzwerte am besten Rechnung getragen wird.
Die verwendete Datenstruktur basiert auf einer Untersuchung von Altmann et al. (2005).

2 Material und Methode

2.1 Modell und Simulationsvarianten

Für unsere Untersuchungen werden Beobachtungen simuliert, die als Realisationen einer Zufallsvariable y angesehen werden, für die das folgende lineare Regressionsmodell gelten soll:

$$y = \beta_1 x_1 + ... + \beta_n x_n + e. \qquad (1)$$

Für die Zufallsvariable y und die Einflussgrößen $x' = (x_1,...,x_n)$ wird mehrdimensionale Normalverteilung vorausgesetzt, wobei gilt:

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N(0,\Sigma) \text{ mit } \Sigma = Var\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma'_{yx} \\ \sigma_{yx} & \Sigma_x \end{pmatrix}.$$

Die in unseren Untersuchungen verwendeten Einflussgrößen $x' = (x_1,...,x_n)$ seien entsprechend der Größe von $\hat{\beta}_i / SE(\hat{\beta}_i)$ bei Nutzung der Ergebnisse von Altmann et al. (2005) absteigend sortiert.
Die Simulation erfolgt bei systematischer Variation der Modellkomplexität. Damit kann untersucht werden, ob mit Hilfe der einbezogenen Kriterien bzw. Selektionsverfahren nicht im Modell befindliche Variable auch erkannt werden. Die Simulationsvarianten von Modell (1) ergeben sich durch Modifikation von Σ.

<u>Variante: Variable 1 bis Variable 1-8</u>
Beginnend mit Variable 1 wurden schrittweise weitere Variablen in das Modell zur Simulation von Datensätzen $(y, x_1,..., x_p, x_{p+1},..., x_n)$ aufgenommen. Hierbei sind $(x_{p+1},..., x_n)$ fiktive Variable ohne Covarianz zu den Variablen im Modell. Im vorliegenden Fall gilt n=8 und p=1, 2,…,8.

$$Var\begin{pmatrix} y \\ x_p \\ x_{n-p} \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx_p}' & 0 \\ \sigma_{yx_p} & \Sigma_{x_p} & 0 \\ 0 & 0 & \Sigma_{x_{n-p}} \end{pmatrix}.$$

Variante: Variable 2-8 bis Variable 8

Beginnend mit Variable 1 wurden schrittweise weitere Variablen aus dem Modell zur Simulation von Datensätzen $(y, x_1,...,x_p, x_{p+1},...,x_n)$ eliminiert. Hierbei sind $(x_1,...,x_p)$ fiktive Variable ohne Covarianz zu den Variablen im Modell, wobei hier n=8 und p=1, 2,…,7 gilt.

$$\mathrm{Var}\begin{pmatrix} y \\ x_p \\ x_{n-p} \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & 0 & \sigma'_{yx_{n-p}} \\ 0 & \Sigma_{x_p} & 0 \\ \sigma_{yx_{n-p}} & 0 & \Sigma_{x_{n-p}} \end{pmatrix}.$$

Entsprechend der jeweils verwendeten Struktur von Σ wurde die Restvarianz der Zielgröße angepasst. Dabei war die Varianz der Zielgröße über alle Fragestellungen und Varianten konstant mit dem Wert 1, um für die Zielvariable über alle Modelle eine vergleichbare Varianz zu sichern.

2.2 Benutzte (Co)Varianzmatrix

Die in den Untersuchungen von Altmann et al. (2005) geschätzte Korrelationsmatrix wurde als wahr angesehen und für die Simulation gemäß Modell (1) verwendet.

Tabelle 1: Verwendete Korrelationsmatrix (nach Altmann et al., 2005)

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7	Variable 8	Zielgröße
Variable 1	1	0.258	0.288	0.001	-0.084	0.325	-0.038	-0.070	0.466
Variable 2		1	0.974	-0.603	-0.598	0.664	-0.641	-0.544	0.298
Variable 3			1	-0.489	-0.626	0.670	-0.597	-0.579	0.328
Variable 4				1	0.724	0.052	0.884	0.582	-0.284
Variable 5					1	0.109	0.945	0.954	-0.356
Variable 6						1	0.101	0.131	0.105
Variable 7							1	0.888	-0.320
Variable 8								1	-0.289
Zielgröße									1

Die Korrelationen zwischen den Einflussgrößen und der Zielgröße liegen zwischen -0.356 und 0.466.

2.3 Verwendete Bewertungskriterien

In unseren Untersuchungen wurden die in SAS (Prog Reg) verfügbaren Bewertungskriterien berücksichtigt (Tabelle 2).

Tabelle 2: Verwendete Bewertungskriterien

Abkürzung, Berechnung, Quelle
$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2p$ (Akaike, 1969)
$AIC\_cor = n \cdot \ln\left(\frac{SSE}{n}\right) + \frac{n(n+p)}{(n-p-2)}$ (Hurvich und Tsai, 1989)
$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + p\ln(n)$ (Schwarz, 1978)
$BIC\_cor = n \cdot \ln\left(\frac{SSE}{n}\right) + 2(p+2)q - 2q^2$ (Sawa, 1978)
$RSQ = 1 - \frac{SSE}{SST}$
$ADJRSQ = 1 - \frac{n\left(-\frac{SSE}{SST}\right)}{n-p}$
$CP = \frac{SSE}{\hat{\sigma}^2} + 2p - n$ (Mallows, 1973)
$JP = \frac{(n+p)}{n(n-p)}SSE$ (Hocking, 1976)
$PC = \frac{(n+p)}{(n-p)}\frac{SSE}{SST}$ (Judge et al., 1980)
$SP = \frac{1}{(n-p-1)}\frac{SSE}{(n-p)}$ (Hocking 1976)
$GMSEP = \frac{(n+1)(n-2)}{n(n-p-1)}\frac{SSE}{(n-p)}$ (Darlington, 1968)

In Tabelle 2 bedeuten:

n = Anzahl Beobachtungen; p = Anzahl Modellparameter

SST = Summe der SQ der abhängigen Variable; SSE = Summe der SQ der Re steffekte;

$MSE = \frac{SSE}{n-p}$; $\hat{\sigma}^2$ = MSE für das Modell mit allen E inf lußgrössen; $q = \frac{n\hat{\sigma}^2}{SSE}$

Alle angeführten Kriterien basieren auf der Nutzung von SSE und einer unterschiedlichen Wichtung von Stichprobenumfang und der Anzahl Parameter im Regressionsmodell.
Weiterhin werden drei Vorgehensweisen zur Modellwahl untersucht, die basierend auf einer F-Statistik die jeweiligen Regressionskoeffizienten auf Signifikanz prüfen.
Dabei werden bei der sog. „backward selection“ schrittweise alle diejenigen Variablen eliminiert, deren Regressionskoeffizienten nicht signifikant sind. Das aufbauende Verfahren (forward selection) baut schrittweise das anzuwendende Regressionsmodell auf. Hier wird zunächst der Regressionskoeffizient mit dem höchsten F-Wert gesucht. Dazu werden im nächsten Schritt weitere Variable mit signifikanten Regressionskoeffizienten ermittelt, bis keine weitere Variable einen signifikanten Koeffizienten aufweist. Einmal im Modell enthaltene Variable bleiben enthalten. Eine Modifikation des aufbauenden Verfahrens stellt die sog. „stepwise selection“ dar. Hier wird bei Hinzufügung einer weiteren Variable überprüft, ob bereits enthaltene Variable noch signifikant sind. Variablen bleiben nur bei Gültigkeit dieser Bedingung im Modell.
Mit der Nutzung der angeführten Selektionsverfahren ist das grundsätzliche Problem der Abhängigkeit der Testentscheidung eines Schrittes vom Ergebnis des vorherigen Schrittes verbunden. Diese Abhängigkeit erlaubt keine Kontrolle des Risikos von Fehlentscheidungen. Da diese Vorgehensweisen jedoch in vielen Softwarepaketen implementiert sind, werden diese Verfahren hier ebenfalls einbezogen.

2.4 Simulation und Berechnung des Vorhersagefehlers

Entsprechend der Darstellung in 2.1 ergeben sich 15 „wahre“ Modellvarianten. Basierend auf diesen Varianten erfolgt die Simulation der Realisationen der Zielvariable y. Für die Auswertung resultiert im vorliegenden Fall aus $\sum_{k=1}^{8}\binom{8}{k}$ eine Anzahl von 255 möglichen unterschiedlichen Modellen. Für jedes dieser Auswertungsmodelle wurden je Simulationslauf die Parameter der Regressionsfunktion geschätzt und die Kriterien entsprechend Tabelle 2 berechnet. Weiterhin wird für jeden Lauf eine Modellwahl nach den in 2.3 angeführten Verfahren „forward selection“, „backward selection“ und „stepwise selection“ vorgenommen.
Bei Nutzung der Parameterschätzungen aus Datenblock 1 und der Einflussvariablen in Datenblock 2 kann ein Vorhersagewert erzeugt werden. Aus der quadrierten Differenz dieses Vorhersagewertes für die simulierte Zielvariable in Datenblock 2 und der Realisation für die Zielvariable in Datenblock 2 wird ein mittlerer Vorhersagefehler (PMSE) berechnet (Abbildung 1). Dieser Vorhersagefehler kann zur Einschätzung der Wirksamkeit der Bewertungskriterien und Selektionsverfahren verwendet werden. Dabei ist das Kriterium bzw. Selektionsverfahren zu bevorzugen, dass die Modelle mit dem geringsten PMSE auswählt.
Für die Datensimulation sowie Parameterschätzung, Kriterienberechnung und Anwendung der Selektionsverfahren wurde die Statistik-Software SAS (Proc IML, Proc

Reg) verwendet. Die Berechnung der PMSE und Datenauswertung wurde mit dem relationalen Datenbanksystem MS-ACCESS durchgeführt.

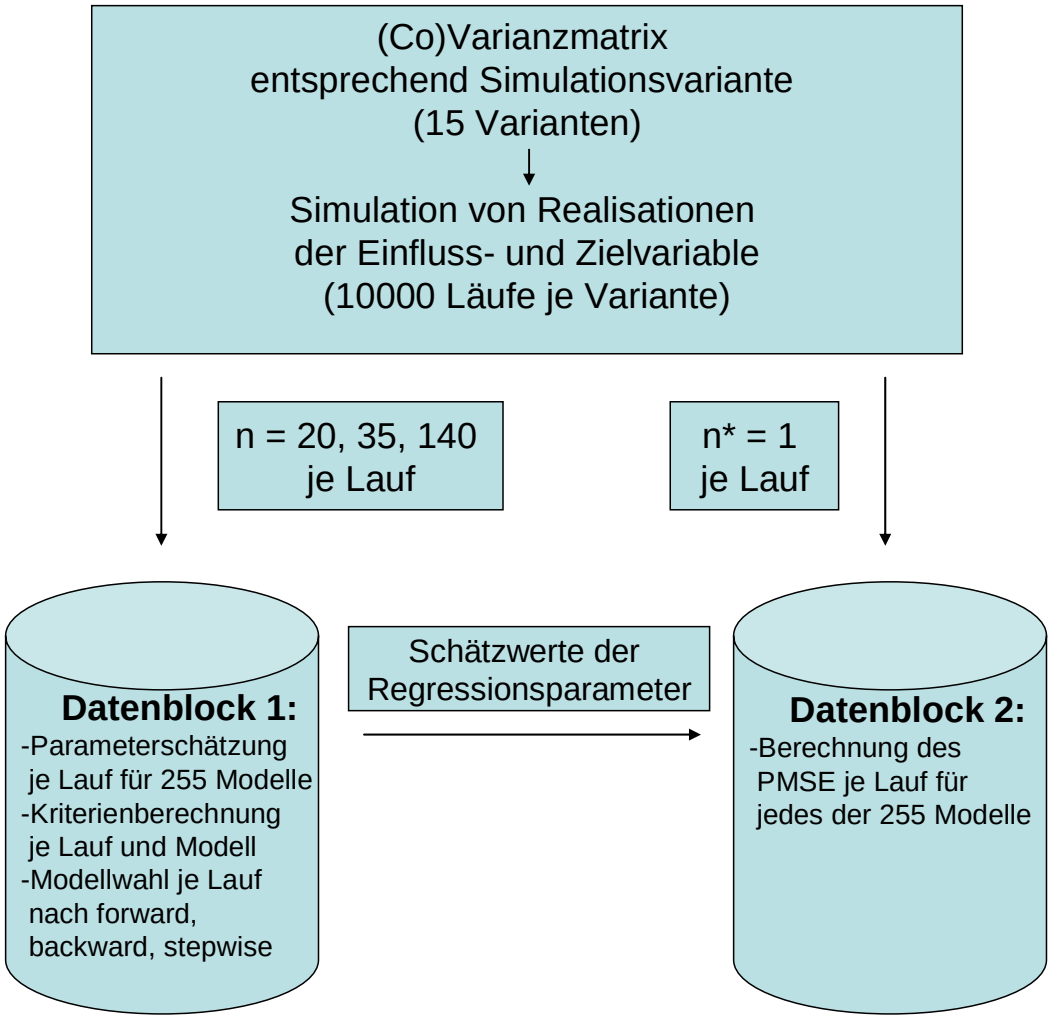

Abbildung 1: Schema der Simulation und Datenauswertung

3 Ergebnisse

Die beschriebene Simulation wird mit drei unterschiedlichen Stichprobenumfängen durchgeführt. Der Umfang n=140 ist von den Untersuchungen von Altmann et al. (2005) abgeleitet. Da bei vielen praktischen Untersuchungen häufig geringere Umfänge vorliegen, wurden weiterhin im Interesse einer Aussageerweiterung die Umfänge n=20 und n=35 einbezogen. Die nachfolgenden Darstellungen zeigen den berechneten Vorhersagefehler (PMSE) bei Nutzung der Bewertungskriterien, bezogen auf AIC_cor. Entsprechend führt ein kleinerer Vorhersagefehler gegenüber AIC_cor zu Werten < 1, ein größerer Vorhersagefehler zu Werten >1.

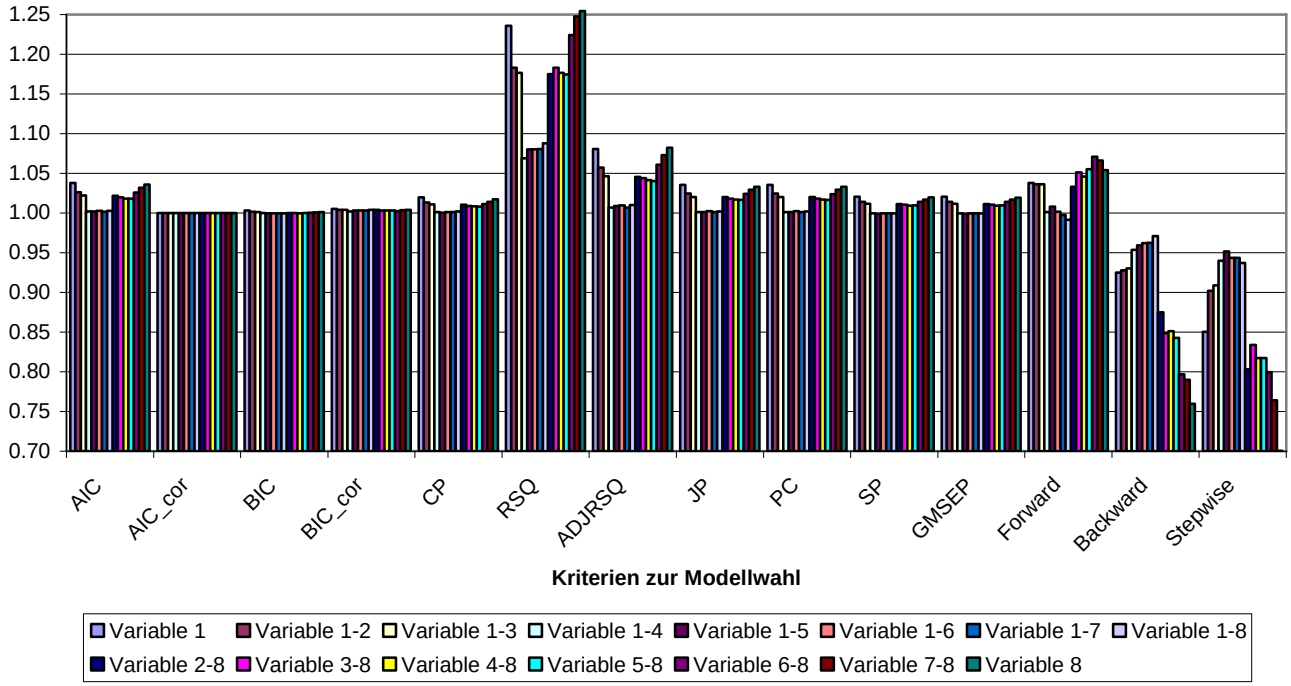

Abbildung 2: PMSE (relativ zu AIC_cor) für die untersuchten Simulationsvarianten, Bewertungskriterien und Selektionsverfahren (n=20)

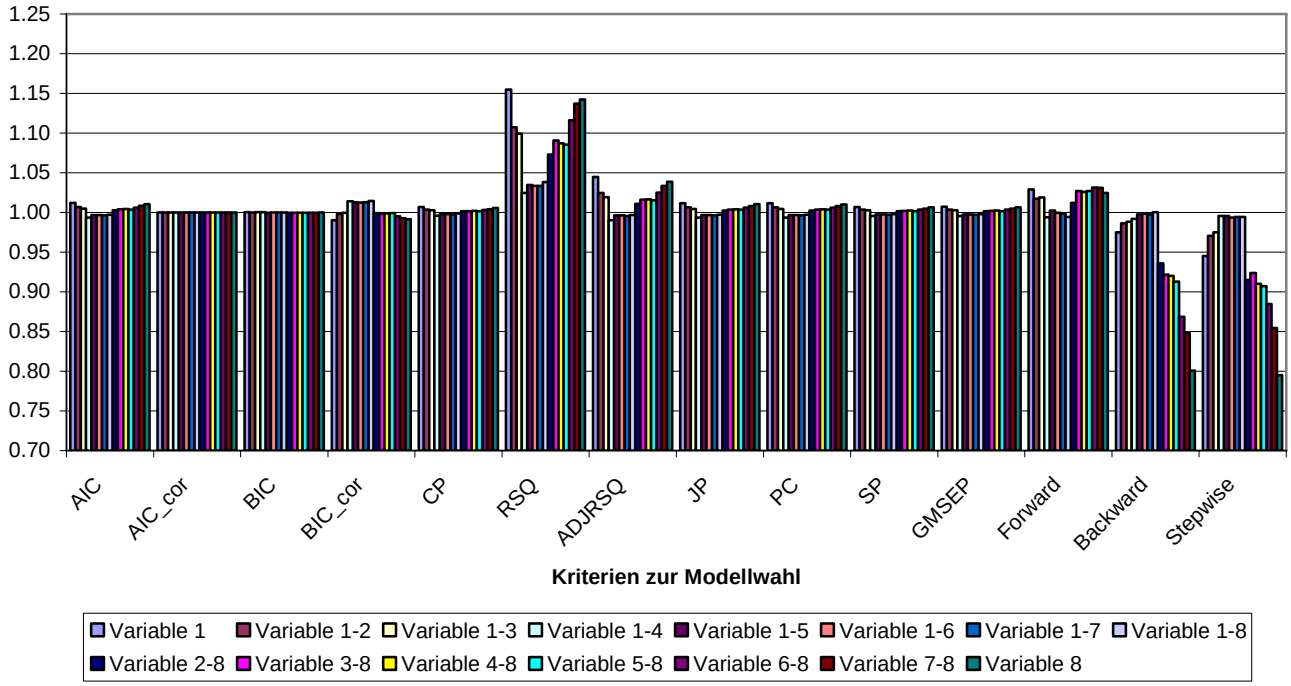

Abbildung 3: PMSE (relativ zu AIC_cor) für die untersuchten Simulationsvarianten, Bewertungskriterien und Selektionsverfahren (n=35)

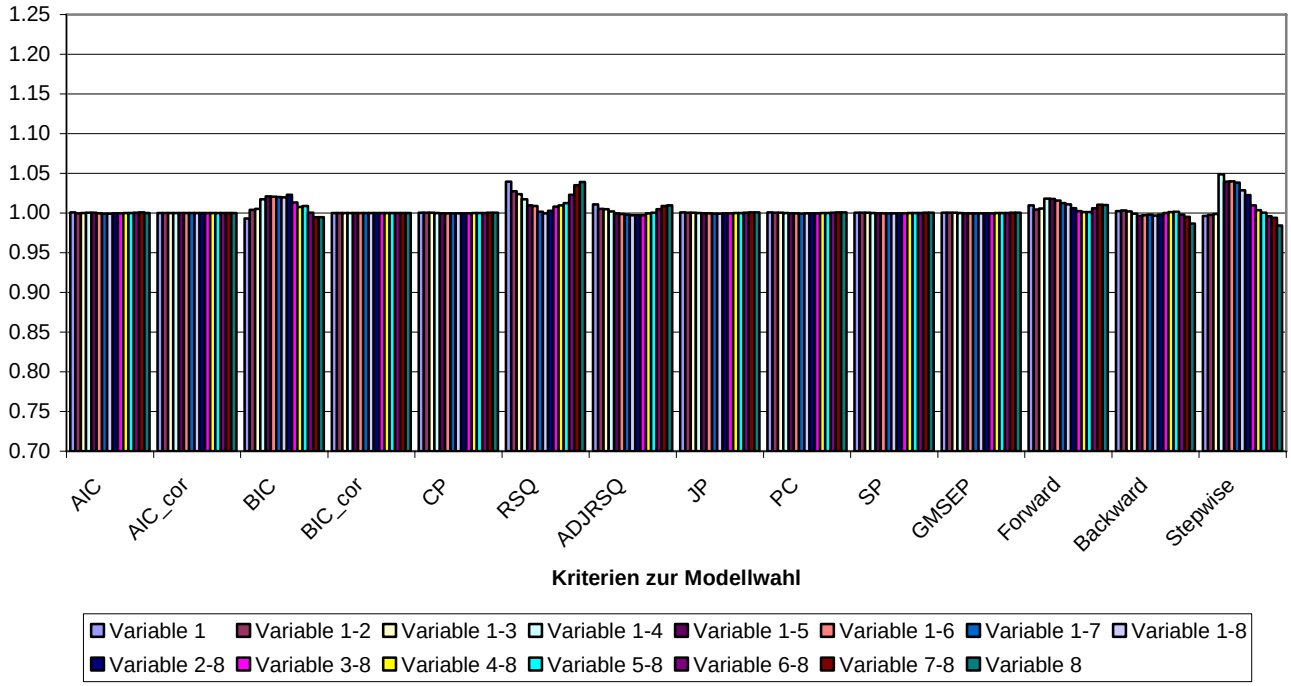

Abbildung 4: PMSE (relativ zu AIC_cor) für die untersuchten Simulationsvarianten, Bewertungskriterien und Selektionsverfahren (n=140)

Die in den Abbildungen 2-4 zusammengestellten Ergebnisse der Vorhersagefehler zeigen eine deutliche Abhängigkeit von den einbezogenen Variablen und vom Stichprobenumfang.
Bei geringem Stichprobenumfang liefern gegenüber AIC_cor von den übrigen analytischen Kriterien nur BIC und BIC_cor sowie von den testbasierten Kriterien „backward selection“ und „stepwise selection“ vergleichbare Ergebnisse. Auffällig ist die Unterlegenheit besonders bei Modellen mit geringer Variablenanzahl, beispielsweise für AIC, RSQ oder JP.
Bei großem Stichprobenumfang ist eine Unterlegenheit nur noch für RSQ, ADJRSQ und „selection forward“ zu beobachten. Weiterhin aber auch für BIC und „selection stepwise“, was bei kleinem Stichprobenumfang nicht der Fall war. Die übrigen Kriterien fallen mit AIC_cor praktisch zusammen.

4 Diskussion und Schlussfolgerungen für die Versuchspraxis

Für die Ergebnisdarstellung wurde das korrigierte AIC (AIC_cor; Hurvich und Tsai, 1989) als Bezugspunkt gewählt. Es zeigt sich, dass für die untersuchte (Co)Varianzstruktur und den Datenumfang kein Bewertungskriterium einen durchgehend geringeren Vorhersagefehler aufweist. Entsprechend wird von den betrachteten Bewertungskriterien AIC_cor zur Anwendung empfohlen. Die Anwendung dieses Kriteriums setzt die Berechnung aller möglichen Modellvarianten voraus. Diese

Forderung stellt aber bei Beachtung der meist verfügbaren Rechentechnik keine Einschränkung für die praktische Anwendung dar.
Die günstigeren Ergebnisse für „backward selection und „stepwise selection“ bei kleinem Datenumfang können bei größerem Datenumfang insbesondere für „stepwise selection“ nicht beobachtet werden. Dennoch stellen diese Vorgehensweisen bei geringem Datenumfang eine mögliche Alternative zu AIC_cor dar.

Literatur

[1] Akaike, H. (1969): Fitting autoregressive models for prediction, Annals of the Institute of Statistical Mathematics, 21, 243-247.

[2] Altmann, M.; Pliquett, U.; Suess, R.; v. Borell, E. (2005): Prediction of carcass composition by impedance spectroscopy in lambs of similar weight. Meat Science 70, 319-327.

[3] Burnham, K.P.; Anderson, D.R. (2004): Multimodel inference. Sociological Methods and Research, 33, 261-304.

[4] Darlington, R.B. (1968): Multiple Regression in Psychological Research and Practice, Psychological Bulletin, 69, 161 - 182.

[5] Draper, N.R.; Smith, H. (1981): Applied regression analysis. Wiley, New York.

[6] Hocking, R.R. (1976): The Analysis and Selection of Variables in Linear Regression, Biometrics, 32, 1 - 50.

[7] Hurvich, C.M.; Tsai, C.L. (1989): Regression and time series model selection in small samples, Biometrika 76, 297-397.

[8] Judge, G.G., Griffiths, W.E., Hill, R.C., and Lee, T. (1980): The theory and practice of econometrics, New York: John Wiley & Sons, Inc.

[9] Mallows, C.L. (1973): Some comments on C_p, Technometrics, 15, 661-675.

[10] Rasch, D.; Herrendörfer, G.; Bock, J.; Victor, N.; Guiard, V. (1996): Verfahrensbibliothek Band I. R. Oldenbourg Verlag München Wien.

[11] Sawa, T. (1978): Information Criteria for Discriminating Among Alternative Regression Models, Econometrica, 46, 1273 - 1282.

[12] Schwarz, G. (1978): Estimating the dimension of a model, Annals of Statistics, 6, 461 – 464.

[13] Wetherill, G.B. (1985): Regression analysis with applications. Chapman and Hall, London, New York.