Randomized Model Selection as a Model Building Strategy in Stepwise Logistic Regression

060

Marcus Vollmer
Department of Mathematics and Computer Science
University of Greifswald, Germany

ERNST MORITZ ARNDT
UNIVERSITÄT GREIFSWALD

Wissen lockt. Seit 1456

Introduction

Logistic regression is an approved method in the applied research for predicting the outcome of a binary variable using predictor variables (see Harrell[2] and Steyerberg[6]). Especially in medical research, logistic regression is a well known method for **predicting the survival of patients or verifying the therapeutic outcome**. If many variables are involved, it can be difficult to select a good set of relevant variables, including interactions.
Many publications of critical and intensive care medicine were analyzed concerning the usage of logistic regression. Most authors (about 50%) use the in-hospital mortality as the response variable, about 25% use 28- or 30-day survival. In some publications, it was not specified, how the multivariate model was generated. Often the multivariate modeling was performed using the statistically significant univariate variables (e.g. Nguyen et al.[3]).

Hierarchical procedures, like the **stepwise logistic regression with backward-elimination and forward-selection are commonly used** (Sebat et al.[4] or Shapiro et al.[5]). The aim of the procedure is to improve an initial model by adding or removing one predictor (1st Generation). The limited number of possible models of the 1st generation are compared using a criterion for model selection, such as the Bayesian information criterion (BIC). For a model with k predictor variables and sample size n it is:

$$BIC = -2 \log L(\hat{\theta}) + k \log n$$

The initial model will be modified towards the 1st-generation-model with the lowest BIC value. The "mutation" ends, if no improvement is possible (see chart A below). Nevertheless, stepwise methods have various disadvantages[6], like instability of the selection and biased estimation of coefficients.

Randomized Model Selection

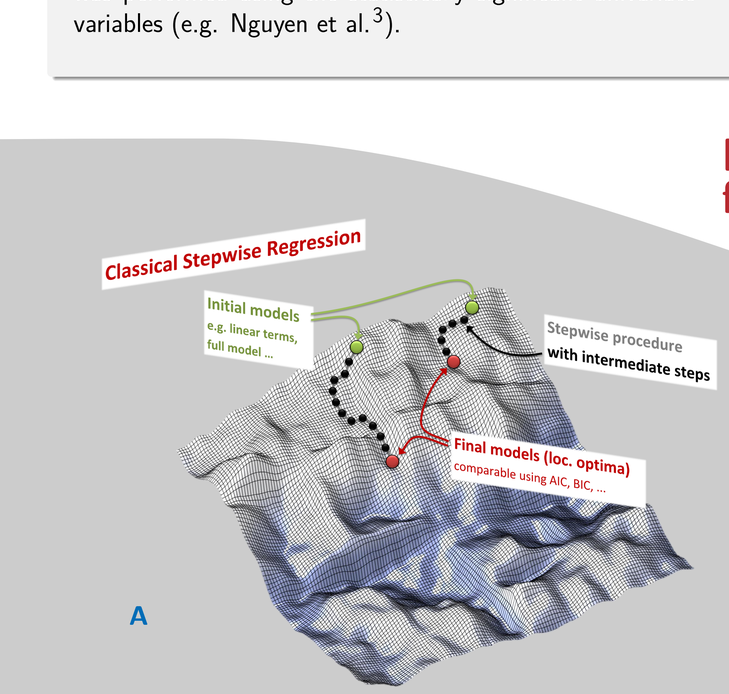
The classical stepwise regression ends only in a local optimum and depends strongly on the initial model (see chart A and B).

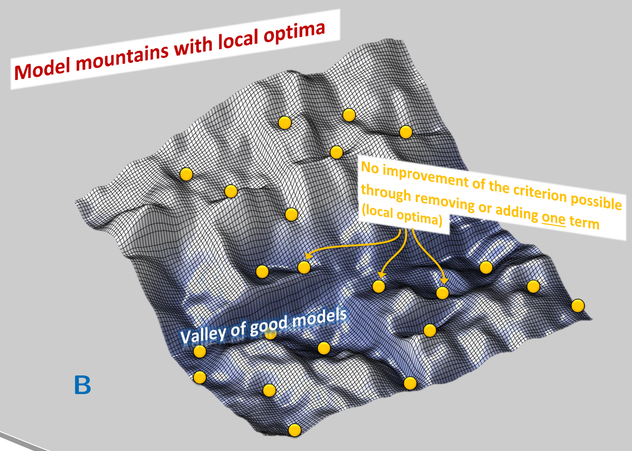
Randomized Model Selection - heuristic technique & modified stepwise procedure

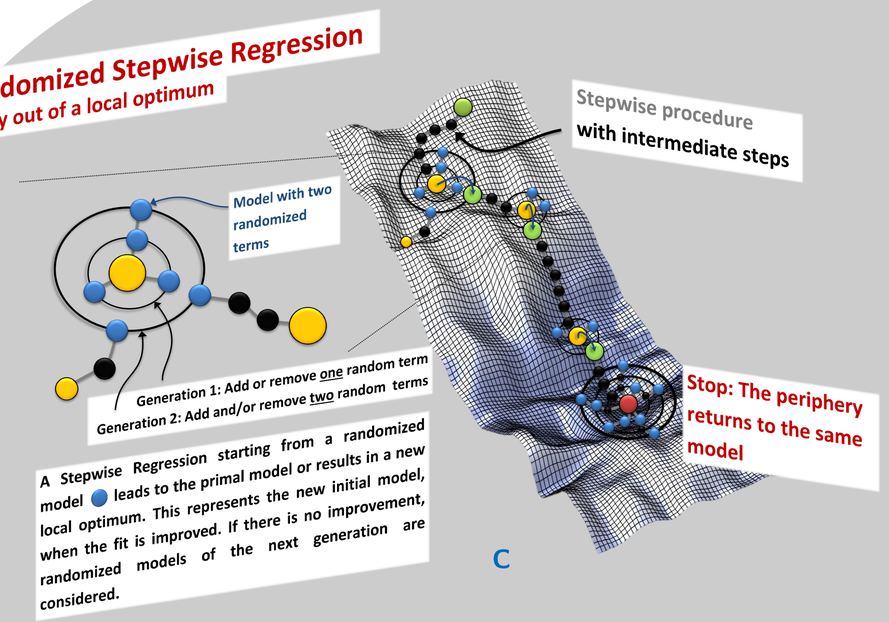
- Initial model randomly chosen
- Classical stepwise procedure tries to optimize the logistic model until a local optimum is reached
- Only g_1 modified models of the first generation are considered
- From each of these g_1 models the classical stepwise procedure tries to optimize
- Either the procedure leads to the previous optimum or it ends in some new local optimum
- The best model, concerning the BIC value, will be selected
- If none of the g_1 final models provides an improvement, g_2 randomly chosen models of the second generation will be considered
- The procedure ends, unless the l-th generation improves the logistic model

The number of generations l, its sizes $g_1, \ldots, g_l$ and the number of involved variables determine the computing time.

Modern heuristic techniques are needed for finding a satisfactory model.







Application to Sepsis data and validation of the new procedure

To rate the performance of the new strategy, we investigate the intensive care unit survival of 746 patients (see Gruendling[1]) with septic shock or severe sepsis between 2006 and midyear 2012. In 2008, a continuous training program of diagnostic measures and treatment was implemented.

Response Variable: 90-day-survival
Possible Predictors:

- Demographic characteristics *(age, gender)*
- Septic diagnosis *(severeness, focus, SAPSII-Score, APACHEII-Score, lactate, location of diagnosis [ICU, IMC/surgery/normal station, Emergency])*
- Therapy *(preexisting antibiotic administration, adequate primary antibiotic administration, time to antibiotic administration, taking smear of the focus, crystalloid administration first 6h and 24h, hydroxyethyl starch first 24h)*
- Training program group

Model: Logistic regression with bivariate interactions

Classical stepwise logistic regression with different initial models were compared with the randomized approach. For the randomized model selection, 20 randomly chosen models were taken. Each contains 10 terms. The method proceeds till the fourth generation ($[g_1, \ldots, g_4] = [5, 10, 15, 20]$). Initial models for the classical stepwise model selection were the empty model, the model with all linear terms, half of all possible interaction terms (Interaction) and three random models including 10, 15 and 20 terms (linear and interaction terms mixed). Both procedures optimize according to the BIC.

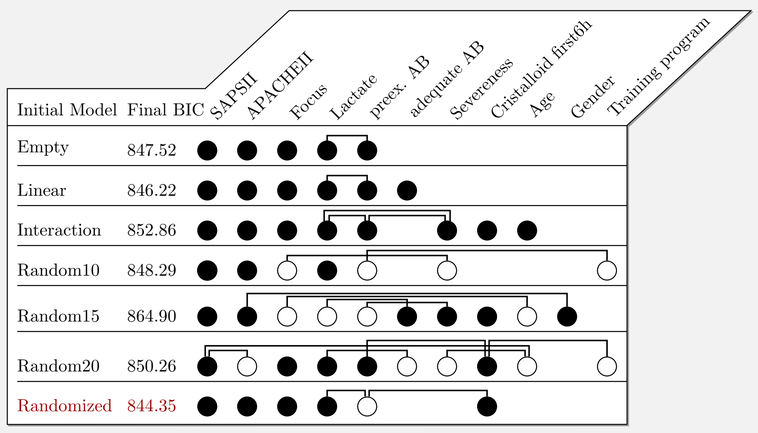


Figure 1 : Predictors of the final models. Filled circles: linear terms. Lines between circles: interactions. Unfilled circles: appearance only in interactions.

Figure 1 shows, which predictors and interactions are involved in the final models. Each final model represents a local optimum and illustrates how different the logistic models can be. **The quality of the models is expressed by the BIC values.**

The best model might be the one from the randomized model selection. It contains five linear terms (SAPSII, APACHEII, focus, lactate, cristalloid administration first 6h) and two interaction terms involving the preexisting antibiotic administration.

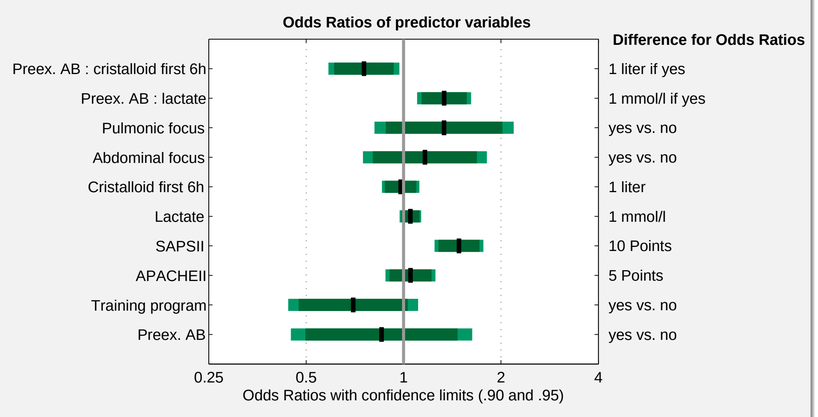
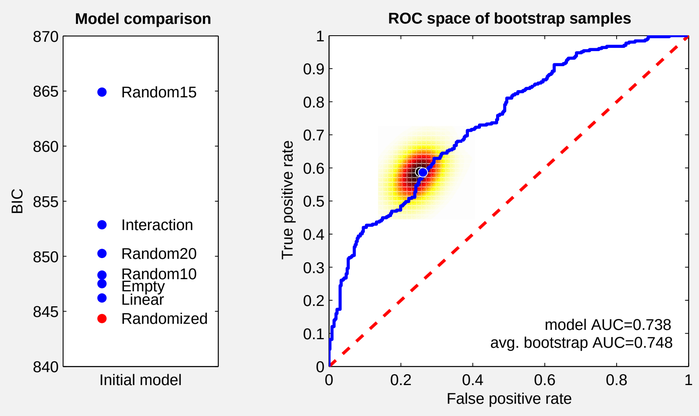


Figure 2 : Odds Ratios of the predictors.

For the better interpretation, the linear terms of the interactions were added to model. In addition the binary variable training program were added to quantify the outcome of the continuous training program. Figure 2 shows the corresponding odds.
Bootstrapping (see Steyerberg[6], Chapter 5.3) was used to evaluate the stability of the regression coefficients and of the AUC value (see figure 3b).



(a) BIC values of the different models.

(b) ROC curve and bootstrapping inference of the best model.

Figure 3 : Validation of the final model.

Conclusions

- ✓ Randomized algorithm was more successful
- ✓ Construction ensures to produce better or equivalent models
- ✓ Model selection seems to be more stable
- ✓ Random initial models can be used for enhancement
- ✓ Success of the training program of Sepsisdialog was verified
- ✓ Cristalloid administration, part of the therapeutic intervention, reduced the mortality.

Briefly

LOGISTIC REGRESSION
improved by random algorithm stable, better models

Predictors of 90d-survival?
patients with septic shock or severe sepsis

Training program with cristalloid administration reduces mortality

References

[1] Gruendling, M. (2013). Outcomeverbesserung bei sepsis - ergebnisse eines kontinuierlichen ausbildungsprogramms. *Der Anaesthesist*. submitted.
[2] Harrell, F. (2001). *Regression Modeling Strategies*. Springer, corrected edition.
[3] Nguyen, H. B., Rivers, E. P., Knoblich, B. P., Jacobsen, G., Muzzin, A., Ressler, J. A., and Tomlanovich, M. C. (2004). Early lactate clearance is associated with improved outcome in severe sepsis and septic shock. *Critical Care Medicine*, 32:1637–1642.
[4] Sebat, F., Musthafa, A. A., Johnson, D., Kramer, A. A., Shoffner, D., Eliason, M., Henry, K., and Spurlock, B. (2007). Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years. *Critical Care Medicine*, 35:2568–2575.
[5] Shapiro, N. I., Howell, M. D., Talmor, D., Lahey, D., Ngo, L., Buras, J., Wolfe, R. E., Weiss, J. W., and Lisbon, A. (2006). Implementation and outcomes of the Multiple Urgent Sepsis Therapies (MUST) protocol. *Critical Care Medicine*, 34:1025–1032.
[6] Steyerberg, E. W. (2008). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health)*. Springer, 1st edition.

Acknowledgments

Special thanks to the team of Sepsisdialog (www.sepsisdialog.de) of the Universitätsmedizin Greifswald, managed by Dr. Sigrun Friesecke and Dr. Matthias Gründling, for providing the dataset and for the good interdisciplinary collaboration.